Neural FOXP2 - Language Specific Steering for Targeted Language Improvement in LLMs

Anonymous ACL submission

Abstract

Large Language Models are multilingual, yet their default generation prior is often English due to pretraining imbalance. We hypothesize that language defaultness is controlled by a sparse, low-rank circuit of language-specific features that can be causally localized and steered.

We introduce **Neural FOXP2**, an inference-time method that shifts a models default language without retraining. First, per-layer sparse autoencoders (SAEs) decompose activations into sparse features, enabling identification of language-selective features using selectivity and logit-mass lift metrics. Second, we characterize language-shift geometry through layer-wise singular value decomposition of English to target activation differences, suggesting towards compact steering subspaces and stable intervention windows. Third, we apply signed sparse activation steering that promotes target-language directions while suppressing English-default attractors.

We evaluate Neural FOXP2 across five open weight large language models across five languages which are Hindi, Spanish, Chinese, Bengali, and Telugu. Across translation, question answering, natural language inference, and summarization, Neural FOXP2 consistently increases target-language defaultness while preserving task semantics, making the target language the default realization channel without full retraining.

1 Introduction

Modern decoder-only LLMs are trained on web-scale corpora with heavily imbalanced language distributions, where high-resource languages such as English dominate token exposure (Hu et al., 2025; Guo et al., 2025). As a result, multilingual models often develop an English-default generation prior, making other languages less likely to emerge under weak or ambiguous prompts (Shani and Basirat, 2025). Existing approaches for shifting this default behavior, including continued pretraining and instruction tuning, can introduce capability drift, cross-lingual interference, and safety degradation due to distributional shift.

Prior signals and Gap Analysis Recent work suggests that multilingual behavior is partially governed by localized internal mechanisms rather than prompts alone. Prior studies show that language-specific units can influence multilingual generation (Tang et al., 2024; Kojima et al., 2024a), multilingual models can exhibit asymmetric language preferences (Shani and Basirat, 2025), and causal intervention methods such as activation patching enable targeted editing of internal representations (Zhang and Nanda, 2023; Meng et al., 2022; Rahmanisa et al., 2025). Recent evidence also suggests that multilingual reasoning and language realization may be partially disentangled, indicating that language control can be modified without fully disrupting reasoning behavior (Zhao et al., 2025).

However, these directions have not been unified into a stable framework for targeted language steering. In particular, existing methods lack (i) a mechanistic objective for measuring early language commitment (Tezuka and Inoue, 2025), (ii) a feature-level localization strategy robust to superposition, and (iii) stability constraints that bound regressions during intervention.

We hypothesize that language defaultness is governed by a sparse set of language-selective internal features whose intervention can predictably shift the models output language while preserving task semantics. This hypothesis is consistent with prior work suggesting that specific behaviors can arise from localized computations rather than uniformly distributed representations (Meng et al., 2022; Zhong et al., 2026). In the multilingual set-

[0]Code and evaluation pipelines are publicly available. Code.

ting, this implies that targeted language adaptation may be achieved through selective representation steering instead of full parameter updates.

Language defaultness: Let $p_\theta(\cdot \mid x)$ denote the next-token distribution, and let $\mathcal{V}_{\text{tgt}}$ and $\mathcal{V}_{\neg\text{hi}}$ denote Hindi and non-Hindi token sets, respectively. We define the early-step language advantage as:

$$\Delta_{\text{tgt}}(x) = \log \sum_{t \in \mathcal{V}_{\text{tgt}}} p_\theta(t \mid x) - \log \sum_{t \in \mathcal{V}_{\neg\text{tgt}}} p_\theta(t \mid x)$$

A positive $\Delta_{\text{hi}}(x)$ indicates that the model assigns greater probability mass to Hindi tokens before generation settles into a language trajectory. Our objective is to increase Δ_{hi} on neutral prompts while preserving task semantics.

Feature-based language control. Direct neuron-level interventions are unreliable because transformer representations are highly superposed, with individual units encoding multiple unrelated features (Elhage et al., 2022). To obtain more interpretable and controllable representations, we use Sparse Autoencoders (SAEs) to decompose activations into sparse feature coordinates (Cunningham et al., 2023). We define language neurons as SAE features that consistently track and causally influence language defaultness.

Causal language features: To identify language-selective features, we measure the causal effect of intervening on a feature z_j using activation patching (Zhang and Nanda, 2023). We define the intervention score:

$$\mathcal{I}_j = \mathbb{E}_{x \sim \mathcal{D}}\Big[\Delta_{\text{tgt}}(x; \text{do}(z_j \leftarrow z_j + \epsilon)) - \Delta_{\text{tgt}}(x)\Big]$$

A feature is considered language-relevant if intervention consistently increases target-language probability mass under matched semantic prompts. We further evaluate necessity through ablation and sufficiency through feature patching, ensuring that identified features causally influence language selection rather than merely correlate with Hindi or Spanish text.

Our contribution: Building on prior work in language-specific neurons, causal intervention, and sparse feature representations (Tang et al., 2024; Kojima et al., 2024a; Zhang and Nanda, 2023; Cunningham et al., 2023), we introduce **Neural FOXP2**, a feature-level steering framework for multilingual control. Neural FOXP2 operates in three stages: (1) localizing language-selective SAE features using causal intervention scores, (2) identifying low-rank steering directions through layer-wise SVD of language-shift activations, and (3) applying sparse signed activation steering to increase target-language defaultness while preserving task semantics. To improve stability, interventions are restricted to localized sparse features and calibrated steering subspaces, reducing unintended degradation from broad activation edits.

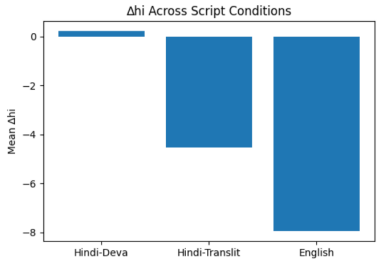

Figure 1: **Hindi language advantage across script conditions.** Devanagari Hindi produces positive Δ_{hi}, while transliterated Hindi and English remain negative, indicating an English-default prior in the base model.

2 Neural FOXP2: Causal Feature Editing for Language Steering

English is the *lingua franca*. Despite multilingual pretraining, multilingual LLMs often exhibit an English-default generation prior. Under weak prompting, Hindi generation typically requires explicit script cues, while transliterated Hindi frequently shifts back toward English **Fig. 1**

We formalize language defaultness using early-step target-language probability mass. Let V_t and V_{en} denote target-language and English token sets, respectively. For decoding context ctx_t with next-token distribution $p_\theta(u \mid \text{ctx}_t)$, we define

$$M_t(x) = \sum_{u \in V_t} p_\theta(u \mid \text{ctx}_t)$$

$$M_{\text{en}}(x) = \sum_{u \in V_{\text{en}}} p_\theta(u \mid \text{ctx}_t)$$

and measure language defaultness as

$$\Delta M(x) = M_t(x) - M_{\text{en}}(x)$$

We evaluate $\Delta M(x)$ over the first decoding steps ($t \in \{1, 2, 3\}$). A successful intervention increases $\mathbb{E}_{x \sim \mathcal{D}}[\Delta M(x)]$ while preserving task semantics.

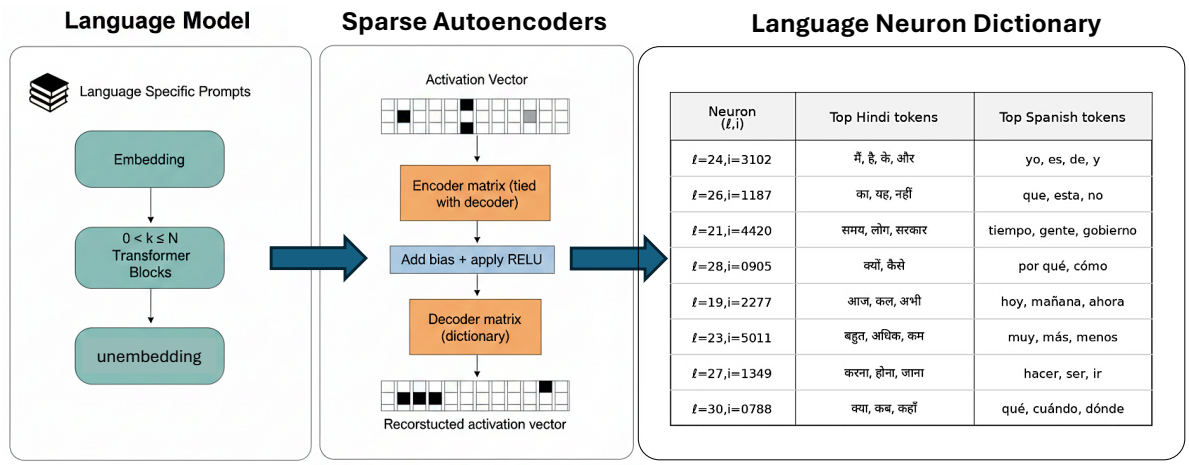

Figure 2: **Neural FOXP2 Stage I overview.** Residual activations from multilingual prompts are decomposed using sparse autoencoders to obtain interpretable dictionary features. Language-selective features are then identified and mapped to Hindi- and Spanish-specific language neurons for downstream steering.

2.1 Stage I: Localize Language Neurons in a SAE Dictionary Basis

Why a SAE dictionary? : Direct neuron-level interventions are unreliable because transformer representations are highly polysemantic, with individual units encoding multiple unrelated features (Elhage et al., 2022). Neural FOXP2 instead defines **language-selective features** in a sparse dictionary basis learned through per-layer SAE autoencoders, yielding feature coordinates that are more interpretable and controllable than raw channels (Bricken et al., 2023; Wu et al., 2026). As shown in **Fig. 2**, activations are decomposed into sparse features whose dominant token affinities form a compact language-neuron dictionary. This is consistent with prior work showing that multilingual behavior can be concentrated in small subsets of units and steered through targeted interventions (Tang et al., 2024; Andrylie et al., 2026).

Activation collection and SAE training: For each layer $\ell \in \{1, \ldots, L\}$, we collect residual-stream activations $h^{(\ell)}(x) \in \mathbb{R}^d$ at the final prompt token. To separate language identity from semantic content, we construct matched prompts $(x_{\text{en}}^{(k)}, x_{\text{tgt}}^{(k)})$ that express the same meaning across languages. The training set $\mathcal{D}_{\text{mix}}$ includes both weakly specified prompts for probing defaultness and explicit-language prompts for estimating language selectivity.

At each layer, we train a sparse autoencoder (SAE) with linear encoder-decoder weights and ReLU activations:

$$z^{(\ell)}(x) = \text{ReLU}\Big((W_\ell)^\top h^{(\ell)}(x) + b_\ell\Big)$$

$$\hat{h}^{(\ell)}(x) = W_\ell z^{(\ell)}(x)$$

where $W_\ell \in \mathbb{R}^{d \times m}$. The SAE is trained using reconstruction loss with an ℓ_1 sparsity penalty:

$$\min_{W_\ell, b_\ell} \mathbb{E}_{x \sim \mathcal{D}_{\text{mix}}} \Big[\|h^{(\ell)}(x) - \hat{h}^{(\ell)}(x)\|_2^2 + \lambda_{\text{sparse}} \|z^{(\ell)}(x)\|_1 \Big]$$

A column-norm constraint on W_ℓ can be used to avoid degenerate scaling. Each coordinate $z_j^{(\ell)}$ defines a sparse dictionary feature, and Stage I identifies a subset of these features as language-selective neurons. .

Defaultness target (early-step mass). Let $V_{\text{tgt}}, V_{\text{en}} \subseteq V$ denote target language and English token sets, respectively. For decoding context ctx_t, we define the target-language mass

$$M_t^{\text{tgt}}(x) = \sum_{u \in V_{\text{tgt}}} p_\theta(u \mid \text{ctx}_t)$$

$$M_t^{\text{en}}(x) = \sum_{u \in V_{\text{en}}} p_\theta(u \mid \text{ctx}_t)$$

We measure defaultness for a target language ℓ_t as:

$$\Delta M^{\ell_t}(x, t) = M_t^{\ell_t}(x) - M_t^{\text{en}}(x),$$

$$t \in \{1, 2, 3\}.$$

Stage I localizes features that are (i) **language-selective** and (ii) **causally increase** ΔM^{ℓ_t} under weak prompting.

2.1.1 Stage I-A: Language Selectivity

Feature activations. For layer ℓ and feature j, the SAE activation is

$$a_j^{(\ell)}(x) = \Big(\mathrm{ReLU}\big((W_\ell)^\top h^{(\ell)}(x) + b_\ell\big)\Big)_j$$

Matched-pair selectivity. For a target language ℓ_t using matched pairs $(x_{\mathrm{en}}^{(k)}, x_{tgt}^{(k)})$, compute

$$\mathrm{Sel}_j^{(\ell,\ell_t)} = \mathbb{E}_k\Big[a_j^{(\ell)}(x_{\ell_t}^{(k)})\Big] - \mathbb{E}_k\Big[a_j^{(\ell)}(x_{\mathrm{en}}^{(k)})\Big]$$

and standardize:

$$\widetilde{\mathrm{Sel}}_j^{(\ell,\ell_t)} = \frac{\mathrm{Sel}_j^{(\ell,\ell_t)}}{\mathrm{Std}_k[a_j^{(\ell)}(x_{\ell_t}^{(k)})] + \mathrm{Std}_k[a_j^{(\ell)}(x_{\mathrm{en}}^{(k)})] + \epsilon}$$

Large positive $\widetilde{\mathrm{Sel}}_j^{(\ell,\ell_t)}$ indicates that feature j activates preferentially for the target language under matched semantics.

2.1.2 Stage I-B: Logit-Mass Causal Lift Toward the Target Token Set

Causal feature intervention To measure causal influence, we intervene on feature (ℓ, j) by applying a small perturbation in dictionary space:

$$z^{(\ell)}(x) \leftarrow z^{(\ell)}(x) + \alpha e_j$$
$$h^{(\ell)}(x) \leftarrow W_\ell\, z^{(\ell)}(x)$$

The modified activations are propagated forward to obtain the perturbed next-token distribution

$$p_{\theta,\alpha}^{(\ell,j)}(\cdot \mid \mathrm{ctx}_t)$$

Logit-mass lift: For target language ℓ_t, define the induced change in defaultness:

$$\begin{aligned}\Delta M_\alpha^{(\ell,j,\ell_t)}(x,t) = &\Big(\sum_{u\in V_{\ell_t}} p_{\theta,\alpha}^{(\ell,j)}(u \mid \mathrm{ctx}_t)\Big)\\ &- \Big(\sum_{u\in V_{\mathrm{en}}} p_{\theta,\alpha}^{(\ell,j)}(u \mid \mathrm{ctx}_t)\Big)\\ &- \Delta M^{\ell_t}(x,t)\end{aligned}$$

Aggregate lift over early steps. We aggregate lift across early decoding steps $T = \{1, 2, 3\}$ over weak prompts $\mathcal{D}_{\mathrm{weak}}$:

$$\mathrm{Lift}_j^{(\ell,\ell_t)} = \mathbb{E}_{x\sim\mathcal{D}_{\mathrm{weak}}}\Big[\frac{1}{|T|}\sum_{t\in T}\Delta M_\alpha^{(\ell,j,\ell_t)}(x,t)\Big]$$

We retain sign and stabilize by estimating a small-perturbation slope over multiple magnitudes:

$$\mathrm{LiftSlope}_j^{(\ell,\ell_t)} = \mathrm{median}_{\alpha\in\{\alpha_1,\alpha_2,\alpha_3\}}\frac{\mathrm{Lift}_j^{(\ell,\ell_t)}(\alpha)}{\alpha}$$

A language neuron must exhibit **positive** $\mathrm{LiftSlope}_j^{(\ell,\ell_t)}$.

2.1.3 Stage I-C: Ranking and the Language-Neuron Set $\mathcal{N}_{\ell_t}$

Feature Ranking For target language ℓ_t, we combine selectivity and causal lift into a joint score:

$$\begin{aligned}S_j^{(\ell,\ell_t)} &= \max\Big(\widetilde{\mathrm{Sel}}_j^{(\ell,\ell_t)}, 0\Big)\\ C_j^{(\ell,\ell_t)} &= \max\Big(\mathrm{LiftSlope}_j^{(\ell,\ell_t)}, 0\Big)\\ \mathrm{Score}_j^{(\ell,\ell_t)} &= S_j^{(\ell,\ell_t)}\cdot C_j^{(\ell,\ell_t)}\end{aligned}$$

High-scoring features are both language-selective and causally effective.

Language feature set For each layer ℓ, we select the top-K features ranked by $\mathrm{Score}_j^{(\ell,\ell_t)}$ to obtain $\mathcal{N}_{\ell_t}^{(\ell)}$. We choose K using lift saturation, adding features until gains in $\mathbb{E}[\Delta M^{\ell_t}]$ plateau. The global language-neuron set is

$$\mathcal{N}_{\ell_t} = \bigcup_{\ell=1}^{L}\mathcal{N}_{\ell_t}^{(\ell)}$$

Language-feature dictionary. For each feature (ℓ, j), we compute a token-affinity signature from its induced pre-softmax logit shift $a_{\ell,j} \in \mathbb{R}^{|V|}$. The top-$k$ target-language tokens with highest affinity form a compact, interpretable signature of the tokens promoted by that feature.

2.2 Stage II: Identify Low-Rank Steering Directions and Choose the Intervention Window

Stage I identifies a compact language-neuron set $\mathcal{N}_{\ell_t}$. Stage II then learns dominant language-shift directions within the same dictionary-feature space and selects a contiguous intervention window $\mathcal{W}$ where these directions are strongest and most stable as illustrated in Figure 3 (Bricken et al., 2023).

Language-Shift activation For target language ℓ_t, we compute SAE feature activations $z^{(\ell)}(x) \in \mathbb{R}^m$ and define the paired language shift

$$\Delta z_k^{(\ell,\ell_t)} \;=\; z^{(\ell)}\big(x_{\ell_t}^{(k)}\big) - z^{(\ell)}\big(x_{\mathrm{en}}^{(k)}\big)$$

To restrict steering to localized language features, we project onto the Stage I neuron support:

$$\Delta\tilde{z}_k^{(\ell,\ell_t)} \;=\; \Pi_{\mathcal{N}_{\ell_t}^{(\ell)}}\,\Delta z_k^{(\ell,\ell_t)}$$

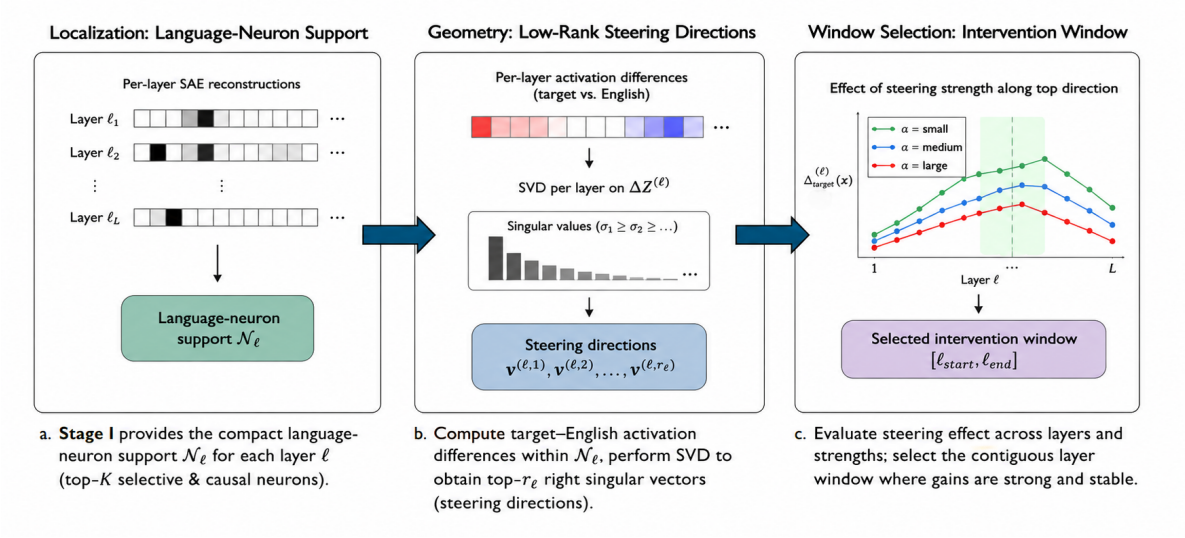

Figure 3: **Neural FOXP2 Stage II overview:** Stage II identifies low-rank steering directions from target–English activation differences using Singular Value Decomposition, and selects a stable intervention window where steering effects are strongest.

We then stack projected shifts into the **language-shift matrix**

$$\Delta Z^{(\ell,\ell_t)} = \begin{bmatrix} (\Delta \tilde{z}_1^{(\ell,\ell_t)})^\top \\ \vdots \\ (\Delta \tilde{z}_N^{(\ell,\ell_t)})^\top \end{bmatrix} \in \mathbb{R}^{N \times m}$$

Layerwise Singular Value Decomposition: We perform a per-layer Singular Value Decomposition:

$$\Delta Z^{(\ell,\ell_t)} = U^{(\ell,\ell_t)} \Sigma^{(\ell,\ell_t)} (V^{(\ell,\ell_t)})^\top$$

where singular values satisfy:

$$\sigma_1^{(\ell,\ell_t)} \geq \sigma_2^{(\ell,\ell_t)} \geq \cdots$$

The right singular vectors $v_i^{(\ell,\ell_t)}$ define dominant steering directions in dictionary-feature space, capturing consistent English to target language shifts across matched prompts. A rapidly decaying spectrum indicates that language steering is governed by a low-rank control structure (Tang et al., 2024; Kojima et al., 2024b).

Steering subspace. For each layer ℓ, we define the steering subspace for the target language

$$\mathcal{S}_{\ell_t}^{(\ell)} = \text{span}\Big(v_1^{(\ell,\ell_t)}, \ldots, v_{r_\ell}^{(\ell,\ell_t)}\Big)$$

Choosing the steering rank. We estimate the steering rank r_ℓ using effective-rank and eigengap diagnostics. Let

$$p_i = \frac{\sigma_i^2}{\sum_j \sigma_j^2}, \qquad r_{\text{eff}}^{(\ell)} = \exp\Big(-\sum_i p_i \log p_i\Big)$$

We initialize

$$r_\ell \leftarrow \left\lceil r_{\text{eff}}^{(\ell)} \right\rceil$$

To identify dominant spectral transitions, we compute

$$g_i^{(\ell)} = \frac{\sigma_i}{\sigma_{i+1}}, \qquad i^\star = \arg\max_i g_i^{(\ell)}$$

The final steering rank is

$$r_\ell = \min\Big(\lceil r_{\text{eff}}^{(\ell)} \rceil, i^\star\Big)$$

Selecting the intervention window $\mathcal{W}$: We score layers by strength and stability:
(i) Spectral Strength: For each layer, we measure the fraction of spectral mass captured by the steering subspace:

$$\text{Mass}_{\ell_t}^{(\ell)} = \frac{\sum_{i=1}^{r_\ell} (\sigma_i^{(\ell,\ell_t)})^2}{\sum_j (\sigma_j^{(\ell,\ell_t)})^2}$$

(ii) Functional sensitivity: We additionally measure the defaultness gain induced by steering:

$$\text{Gain}_{\ell_t}^{(\ell)} = \mathbb{E}_{x \sim \mathcal{D}_{\text{weak}}}\left[\frac{1}{\eta |T|} \sum_{t \in T} \Big(\Delta M_\eta^{\ell_t}(x,t) - \Delta M^{\ell_t}(x,t)\Big)\right],$$

$$T = \{1, 2, 3\}$$

Stability: We evaluate stability by bootstrapping matched prompts, recomputing steering subspaces, and measuring projector overlap:

$$\text{Stab}_{\ell_t}^{(\ell)} = \text{median}_{b \neq b'} \frac{\text{tr}(P_b P_{b'})}{r_\ell} \in [0, 1]$$

Higher values indicate more reproducible steering directions across bootstrap samples.

Choosing the contiguous window: We select a contiguous low-to-mid layer window (Lopo et al., 2025)by maximizing

$$\mathcal{W} = \arg\max_{W} \sum_{\ell \in W} \left(\mathrm{Mass}_{\ell_t}^{(\ell)} \cdot \mathrm{Stab}_{\ell_t}^{(\ell)} \right)$$

with ties resolved using higher gain scores.

Stage II outputs. Stage II returns steering subspaces $\mathcal{S}_{\ell_t}^{(\ell)}$ and an intervention window $\mathcal{W}$, defining the only layers edited during Stage III.

2.3 Stage III: Signed Sparse Steering Targeted to Language Neurons

Intervention-time steering: Stage III applies a sparse inference-time intervention that (i) edits only localized language features, (ii) promotes target-language directions while suppressing English-default activations, and (iii) operates only within the intervention window $\mathcal{W}$. This follows the paradigm of activation steering methods that control model behavior through targeted modifications of intermediate representations (Turner et al., 2024; Panickssery et al., 2023; Zhang et al., 2026).

Let $\Pi_{\mathcal{N}_{\ell_t}^{(\ell)}}$ denote projection onto the localized language-neuron coordinates, and let $P_{\ell_t}^{(\ell)}$ denote the projector onto the steering subspace $\mathcal{S}_{\ell_t}^{(\ell)}$.

Single-layer edit rule : For layers $\ell \in \mathcal{W}$ and a prompt x, we compute the SAE feature representation

$$z^{(\ell)}(x) = \mathrm{ReLU}\big((W_\ell)^\top h^{(\ell)}(x) + b_\ell\big) \in \mathbb{R}^m$$

and then apply a **sparse, feature-gated shift**:

$$\boxed{\begin{aligned} z^{(\ell)}(x) &\leftarrow z^{(\ell)}(x) + \delta z_{\ell_t}^{(\ell)}(x), \\ \delta z_{\ell_t}^{(\ell)}(x) &= \Pi_{\mathcal{N}_{\ell_t}^{(\ell)}}\left(\delta z_{\ell_t}^{(\ell),+}(x) + \delta z_{\ell_t}^{(\ell),-}(x)\right) \end{aligned}}$$

This edit is **feature-gated** by $\Pi_{\mathcal{N}_{\ell_t}^{(\ell)}}$, ensuring we **never perturb** outside the discovered language-neuron coordinates, consistent with sparse-dictionary interpretability goals (Cunningham et al., 2023).

(a) Positive steering component: We define the target steering prototype as the mean English to target activation shift in dictionary space:

$$\mu_{\ell_t}^{(\ell)} = \mathbb{E}_k\left[\Pi_{\mathcal{N}_{\ell_t}^{(\ell)}}\left(z^{(\ell)}(x_{\ell_t}^{(k)}) - z^{(\ell)}(x_{\mathrm{en}}^{(k)})\right)\right]$$

The positive steering update projects this shift onto the low-rank steering subspace:

$$\boxed{\delta z_{\ell_t}^{(\ell),+}(x) = \lambda_\ell \, P_{\ell_t}^{(\ell)} \, \mu_{\ell_t}^{(\ell)}, \; \ell \in \mathcal{W}.}$$

Intuitively: $P_{\ell_t}^{(\ell)} \mu_{\ell_t}^{(\ell)}$ isolates the **most stable, cross-meaning-consistent** directions that govern language change, while λ_ℓ controls intensity (a tunable, small layerwise scale).

(b) Negative steering component: To suppress English-default behavior, we estimate an English attractor direction using weak prompts:

$$\mu_{\mathrm{en}}^{(\ell)} = \mathbb{E}_{x \sim \mathcal{D}_{\mathrm{weak}}}\left[\Pi_{\mathcal{N}_{\ell_t}^{(\ell)}} z^{(\ell)}(x)\right]$$

We then remove the projection of the current feature state onto this attractor:

$$\boxed{\begin{aligned} \delta z_{\ell_t}^{(\ell),-}(x) = -\, \beta_\ell \frac{\langle z^{(\ell)}(x), \mu_{\mathrm{en}}^{(\ell)} \rangle}{\|\mu_{\mathrm{en}}^{(\ell)}\|_2^2 + \epsilon} \mu_{\mathrm{en}}^{(\ell)} \\ \ell \in \mathcal{W}. \end{aligned}}$$

This reduces English-default activation while preserving sparse, localized steering in dictionary-feature space.

Reconstruction and forward propagation: After applying the feature-space intervention, we reconstruct the residual representation:

$$h^{(\ell)}(x) \leftarrow W_\ell \, z^{(\ell)}(x)$$

and continue the forward pass normally (Turner et al., 2024; Panickssery et al., 2023).

Choosing steering magnitude: We tune $(\lambda_\ell, \beta_\ell)$ on a held-out development set using a small grid search and select the largest setting that preserves semantic quality while improving target-language defaultness.

Defaultness gain. We measure the average increase in early-step target-language preference:

$$\Delta_{\mathrm{gain}} = \mathbb{E}_{x \sim \mathcal{D}_{\mathrm{weak}}}\left[\frac{1}{|T|} \sum_{t \in T} \Delta M_{\mathrm{edit}}^{\ell_t}(x,t) - \Delta M^{\ell_t}(x,t)\right]$$

$$T = \{1, 2, 3\}$$

Semantic preservation. We preserve task semantics by enforcing minimal degradation in downstream task performance and limiting distributional

Table 1: **Neural FOXP2: Mechanism Validation.** Each cell reports the delta and absolute mean (baseline→method). We report token-mass shift Δ_{mass}, LID shift Δ_{lid}, and the composed decision score (DEFAULTHI/DEFAULTES). Regression constraints include semantic invariance and guardrails (Spanish bound + KL trust region) (Kojima et al., 2024a). We additionally include a reproduction of LAPE-style neuron activation steering as a baseline (Tang et al., 2024).

		Hindi defaultness				**Spanish defaultness**				**Leakage & stability**		**Utility**
Method	**Edits**	$\Delta_{\text{mass}}\uparrow$	$\Delta_{\text{lid}}\uparrow$	DEFAULTHI↑	$\Delta_{\text{gain}}\uparrow$	$\Delta_{\text{mass}}\uparrow$	$\Delta_{\text{lid}}\uparrow$	DEFAULTES↑	$\Delta_{\text{gain}}\uparrow$	Hi→Es↓	Boot. Stab↑	Task ΔS ↑
No edit	–	+0.00 (−0.55→−0.55)	+0.00 (0.12→0.12)	+0.00 (0.10→0.10)	+0.00 (0.00→0.00)	+0.00 (−0.52→−0.52)	+0.00 (0.14→0.14)	+0.00 (0.11→0.11)	+0.00 (0.00→0.00)	+0.00 (0.00→0.00)	+0.00 (0.00→0.00)	+0.00 (0.00→0.00)
Prompt-only ("Answer in hi/es")	prompt	+0.32 (−0.55→−0.23)	+0.24 (0.12→0.36)	+0.28 (0.10→0.38)	+0.29 (0.00→0.29)	+0.30 (−0.52→−0.22)	+0.21 (0.14→0.35)	+0.26 (0.11→0.37)	+0.27 (0.00→0.27)	+0.10 (0.00→0.10)	+0.18 (0.00→0.18)	−0.06 (0.00→−0.06)
Random feature edits ($\lvert\mathcal{N}\rvert$-matched)	rand-$\mathcal{N}$	+0.06 (−0.55→−0.49)	+0.04 (0.12→0.16)	+0.05 (0.10→0.15)	+0.05 (0.00→0.05)	+0.05 (−0.52→−0.47)	+0.03 (0.14→0.17)	+0.04 (0.11→0.15)	+0.04 (0.00→0.04)	+0.08 (0.00→0.08)	+0.22 (0.00→0.22)	−0.01 (0.00→−0.01)
Edit outside window ($\ell \notin \mathcal{W}$)	$\neg\mathcal{W}$	+0.21 (−0.55→−0.34)	+0.16 (0.12→0.28)	+0.18 (0.10→0.28)	+0.17 (0.00→0.17)	+0.19 (−0.52→−0.33)	+0.14 (0.14→0.28)	+0.16 (0.11→0.27)	+0.15 (0.00→0.15)	+0.12 (0.00→0.12)	+0.41 (0.00→0.41)	−0.02 (0.00→−0.02)
Sparse-only (use $\mathcal{N}_{\ell_t}$, no low-rank $\mathcal{S}$)	$\mathcal{N}$ only	+0.54 (−0.55→−0.01)	+0.46 (0.12→0.58)	+0.50 (0.10→0.60)	+0.49 (0.00→0.49)	+0.52 (−0.52→0.00)	+0.41 (0.14→0.55)	+0.46 (0.11→0.57)	+0.45 (0.00→0.45)	+0.09 (0.00→0.09)	+0.63 (0.00→0.63)	−0.02 (0.00→−0.02)
Low-rank-only (use $\mathcal{S}$, dense over features)	$\mathcal{S}$ only	+0.48 (−0.55→−0.07)	+0.44 (0.12→0.56)	+0.44 (0.10→0.54)	+0.42 (0.00→0.42)	+0.46 (−0.52→−0.06)	+0.39 (0.14→0.53)	+0.41 (0.11→0.52)	+0.40 (0.00→0.40)	+0.18 (0.00→0.18)	+0.52 (0.00→0.52)	−0.03 (0.00→−0.03)
Neural FOXP2 (full): $\mathcal{N}_{\ell_t}+\mathcal{S}^{(\ell)}_{\ell_t}+\mathcal{W}$	**FOXP2**	**+0.85** (−0.55→0.30)	**+0.69** (0.12→0.81)	**+0.68** (0.10→0.78)	**+0.88** (0.00→0.88)	**+0.82** (−0.52→0.30)	**+0.66** (0.14→0.80)	**+0.67** (0.11→0.78)	**+0.84** (0.00→0.84)	**+0.03** (0.00→0.03)	**+0.91** (0.00→0.91)	**−0.01** (0.00→−0.01)

drift during early decoding, following trust-region style constraints (Schulman et al., 2017).

Control intensity. At inference time, we expose a single steering parameter $\gamma \geq 0$ that scales the intervention strength:

$$(\lambda_\ell, \beta_\ell) \leftarrow \gamma(\lambda_\ell, \beta_\ell)$$

This provides a continuous trade-off between no steering ($\gamma = 0$) and stronger target-language defaultness ($\gamma > 1$).

Mechanism validation: **Table 1** summarizes the mechanism validation results and ablation studies. We perform several intervention-based controls to verify that the observed gains arise from localized language steering rather than generic perturbation.

(1) Random-feature controls: We replace $\mathcal{N}^{(\ell)}_{\ell_t}$ with random supports of equal size and apply the same signed intervention. These edits produce weak and unstable defaultness gains, indicating that the effect depends on the discovered language-neuron coordinates (Cunningham et al., 2023).

(2) Support ablation: We ablate localized coordinates by masking $z^{(\ell)}_j \leftarrow 0$ for $j \in \mathcal{N}^{(\ell)}_{\ell_t}$. Defaultness gains collapse under this intervention, demonstrating the necessity of the localized language-neuron support.

(3) Window controls: We compare the selected intervention window $\mathcal{W}$ against matched-size windows outside the chosen layer range. Alternative windows exhibit lower stability and weaker defaultness control, supporting the Stage II steering geometry analysis.

(4) Cross-language leakage: When steering toward Hindi, we measure Spanish token mass as a non-target probe (and vice versa). Effective steering increases the target-language mass without inducing comparable gains in other languages, suggesting partially separable language-associated steering features (Kojima et al., 2024b).

(5) Semantic preservation: On matched prompts, we verify that the intervention changes the language of realization more than the underlying semantic content, with bounded degradation in downstream task quality.

Stage III outputs: Neural FOXP2 produces (i) a sparse language-neuron set $\mathcal{N}_{\ell_t}$, (ii) low-rank steering subspaces $\mathcal{S}^{(\ell)}_{\ell_t}$ and an intervention window $\mathcal{W}$, and (iii) a signed sparse inference-time steering rule that shifts the model toward the target language without retraining (Turner et al., 2024; Panickssery et al., 2023; Kojima et al., 2024b).

3 Performance and Experiments

What we measure. We evaluate two complementary properties: *language defaultness* and *utility preservation*. Defaultness measures whether the model naturally commits to the target language under weak prompting, while utility preservation measures whether steering maintains semantic correctness and downstream task performance.

All evaluations use the same neutral-plus-task mixture with a short greedy or teacher-forced prefix ($m = 8$ by default), without transliteration. We additionally enforce semantic and regression guardrails, including cross-language leakage

Model	XNLI Accuracy					QA F1					XL-Sum ROUGE-L					Translation BLEU				
	Hi	Zh	Bn	Te	Es	Hi	Zh	Bn	Te	Es	Hi	Zh	Bn	Te	Es	Hi	Zh	Bn	Te	Es
Llama-2-7B Base	55.4	58.1	49.2	47.0	63.5	62.0	68.4	52.5	50.8	72.1	33.1	37.2	27.0	24.8	41.3	24.5	31.2	19.6	18.1	35.8
Llama-2-7B Steered	**59.1**	**61.4**	**53.8**	**52.1**	**66.0**	**65.5**	**71.0**	**56.7**	**55.0**	**74.0**	**35.8**	**39.0**	**29.8**	**27.5**	**42.9**	**27.8**	**33.8**	**22.9**	**21.7**	**37.6**
Llama-3.1-8B Base	61.2	64.8	55.6	54.1	68.3	68.5	74.2	58.1	56.5	77.0	37.5	41.6	30.9	28.3	44.1	29.3	36.6	23.4	21.5	39.7
Llama-3.1-8B Steered	**64.6**	**66.9**	**60.1**	**58.9**	**70.1**	**71.6**	**76.1**	**61.9**	**60.4**	**78.5**	**39.8**	**43.1**	**33.4**	**31.0**	**45.3**	**32.1**	**38.6**	**26.5**	**24.8**	**41.2**
Qwen-2.5-7B Base	68.5	78.2	61.2	58.5	75.0	74.8	83.0	65.8	63.0	79.9	41.7	48.0	35.1	33.1	45.7	34.2	43.4	27.3	25.6	40.9
Qwen-2.5-7B Steered	**71.1**	**79.4**	**64.5**	**62.1**	**76.5**	**77.2**	**84.1**	**68.9**	**66.4**	**81.3**	**43.5**	**48.9**	**37.2**	**35.5**	**46.8**	**36.4**	**44.5**	**30.1**	**28.5**	**42.1**
Falcon-40B Base	65.0	75.0	60.0	56.0	72.0	70.0	80.0	60.0	58.0	78.0	38.0	45.0	32.0	30.0	44.5	30.6	35.6	24.1	22.3	39.6
Falcon-40B Steered	**68.0**	**77.0**	**63.5**	**60.0**	**74.0**	**73.2**	**81.9**	**63.2**	**62.0**	**80.1**	**40.1**	**46.4**	**34.3**	**32.5**	**45.8**	**33.2**	**37.5**	**27.1**	**25.5**	**41.0**
Mistral-8B Base	63.6	68.0	54.5	52.6	72.6	70.1	73.4	58.8	57.0	78.3	38.3	41.2	31.7	29.6	44.6	31.5	34.0	23.5	21.0	40.0
Mistral-8B Steered	**66.8**	**70.2**	**58.5**	**56.9**	**74.3**	**73.1**	**75.4**	**62.4**	**60.8**	**79.8**	**40.5**	**42.8**	**34.1**	**32.2**	**45.8**	**34.1**	**35.9**	**26.5**	**24.2**	**42.0**

Table 2: Multilingual evaluation across XNLI, QA, summarization, and translation tasks. Steering consistently improves performance, with the largest gains observed in lower-resource Indic languages such as Bengali and Telugu.

checks and KL-based trust-region constraints, ensuring that gains reflect genuine shifts in multilingual defaultness rather than degraded generation quality.

Main comparison. **Table 2** reports multilingual evaluation on XNLI (Conneau et al., 2018), XQuAD question answering (Artetxe et al., 2020), XL-Sum (Narayan et al., 2018) summarization, and FLORES-200 translation (Team et al., 2022). We compare the base model and FOXP2-steered variants across Hindi, Chinese, Bengali, Telugu, and Spanish. These languages span different resource settings, from high-resource languages such as Chinese and Spanish to lower-resource languages such as Bengali and Telugu, which are more prone to English fallback. This setup evaluates whether FOXP2 yields the largest gains where multilingual defaultness is weakest.

Table 2 shows consistent gains across models, tasks, and languages. FOXP2 improves multilingual performance without retraining, with the largest gains appearing in Bengali and Telugu, where English fallback is more common. The method also reduces language drift in open-ended generation tasks such as summarization and translation. Hindi shows smaller but consistent improvements, reflecting stronger pretraining coverage but persistent script-level interference.

Chinese and Spanish show smaller but consistent gains due to their stronger baselines. The trend remains stable across all model families, including Llama, Qwen, Falcon, and Mistral, suggesting that FOXP2 is architecture-agnostic. Improvements are especially visible in QA, summarization, and translation, where decoding drift is more pronounced. Overall, the results support the claim that FOXP2 improves multilingual defaultness while preserving semantic utility.

Interpretation checklist. A successful FOXP2 operating point should increase both Δ_{mass} and Δ_{lid}, thereby improving DEFAULT_TARGETLANG, while preserving stability and semantic fidelity. We additionally require low cross-language leakage, stable behavior under bootstrap evaluation, and bounded distributional drift under KL trust-region constraints. Together, these checks ensure that improvements reflect genuine shifts in multilingual defaultness rather than unstable or semantically corrupted generation.

4 Conclusion

Neural FOXP2 performs multilingual adaptation through inference-time feature-space steering rather than parameter retraining. We identify sparse language-associated activation features, recover low-rank steering directions, and apply signed sparse activation edits to shift generation toward a target language. Across multiple models and languages, Neural FOXP2 consistently increases target-language defaultness while preserving downstream task performance.

Our results suggest that multilingual behavior can be influenced through sparse activation features rather than uniformly distributed representations. The framework generalizes across Llama, Qwen, Mistral, and Falcon, and remains effective across diverse languages including Hindi, Chinese, Bengali, Telugu, and Spanish. Overall, Neural FOXP2 demonstrates that inference-time activation steering provides a lightweight alternative to multilingual retraining for controllable multilingual generation.

Limitations

Neural FOXP2 is an inference-time steering method that reweights language-relevant representations already present in pretrained models rather than introducing new linguistic competence. Consequently, its behavior depends on the intervention protocol, evaluation metric, and transfer setting.

Construct validity. We measure language defaultness using the early-token signal

$$\Delta M^{\ell_t}(x,t) = M_t^{\ell_t}(x) - M_t^{\mathrm{en}}(x),$$

which captures the models initial target-language preference during decoding. However, ΔM remains a proxy and depends on imperfect token-set definitions, particularly under shared subwords, punctuation, transliteration, and code-mixing. Learned language-ID signals reduce some tokenization artifacts but introduce detector dependence and instability under short prefixes. Most importantly, defaultness is not equivalent to proficiency: entering a target-language mode does not guarantee grammatical quality, factuality, stylistic appropriateness, or downstream utility. Accordingly, defaultness gains should be interpreted together with semantic consistency and downstream task evaluations.

Internal validity. Although the intervention is sparse in dictionary-feature space, it is not perfectly isolated. Transformer representations remain distributed, so feature-space edits can indirectly influence style, verbosity, safety behavior, or other latent functions during downstream computation. In addition, later layers may amplify or attenuate injected components, and local linear approximations may fail for larger intervention magnitudes. Accordingly, our claims are limited to the reported intervention setting: Neural FOXP2 demonstrates measurable effects on language defaultness, but does not imply complete disentanglement of a single language factor. We therefore evaluate stability through ablations, intervention-window controls, saturation tests, and monotonicity analyses across intervention strengths.

Generalization. The method assumes that useful sparse feature representations and low-rank steering directions exist for the model and language under study. These properties may vary across architectures, tokenizers, model scales, and training mixtures, so steering directions and intervention windows may not transfer directly across models without re-validation. Performance may also differ for low-resource languages, underrepresented scripts, heavy code-mixing, or domains with strong stylistic constraints such as safety refusals or long-form reasoning. Consequently, results on controlled prompts should be interpreted as evidence of a controllable multilingual steering signal rather than a guarantee of universal deployment behavior.

Robustness and reproducibility. FOXP2 depends on several design choices, including the language-feature support $\mathcal{N}_{\ell_t}$, steering subspace $\mathcal{S}_{\ell_t}^{(\ell)}$, intervention window $\mathcal{W}$, rank r_ℓ, edit magnitude, token sets, and prompt templates. While these choices make the method inspectable, they also introduce sensitivity to hyperparameters and evaluation settings. Robust evaluation should therefore vary r_ℓ, perturb intervention windows, modify token-set definitions, compare Δ_{mass} and Δ_{lid}, and report multiple random seeds for dictionary training and decoding. Although lighter than retraining, the pipeline still requires activation collection, sparse dictionary learning, low-rank decomposition, and repeated constrained forward passes, making detailed implementation and hyperparameter reporting necessary for reproducibility.

Leakage, safety, and misuse. Language steering may introduce unintended effects such as code-mixing, transliteration drift, semantic changes, or shifts in moderation behavior across languages. We therefore recommend reporting semantic consistency and multilingual safety evaluations alongside defaultness metrics. Although trust-region constraints reduce broad distributional drift, they do not guarantee robustness under adversarial prompting or deployment settings. Consequently, deployment should remain limited to audited languages, bounded intervention strengths, and explicit multilingual red-teaming.

Scope of claims. We do not claim that FOXP2 creates new linguistic competence, replaces multilingual training, perfectly disentangles language from semantics, transfers unchanged across architectures, or guarantees improved user experience from higher defaultness alone. Future work should evaluate additional languages and scripts, compare alternative dictionary designs, strengthen semantic-invariance evaluation, and jointly optimize for defaultness, leakage control, and multilingual safety.

5 Ethics Considerations

Neural FOXP2 is designed to improve multilingual accessibility by enabling controllable language steering without retraining large language models. Our goal is to reduce the dominance of high-resource languages and improve multilingual generation for underrepresented or under-served languages. At the same time, inference-time activation steering introduces potential risks. Modifying language preferences could be misused to manipulate generation behavior, suppress languages, amplify targeted narratives, or bypass language-specific moderation systems. To reduce unintended behavioral drift, Neural FOXP2 incorporates constrained intervention mechanisms including sparse feature-space edits, restricted intervention windows, semantic-preservation checks, trust-region constraints, and cross-language leakage evaluation. However, these safeguards do not guarantee robustness under adversarial prompting or real-world deployment conditions.

Neural FOXP2 does not create new linguistic competence; instead, it reweights multilingual capabilities already present within pretrained models and therefore inherits the biases and limitations of the underlying training data and foundation models. Increased target-language defaultness does not guarantee factual correctness, fairness, cultural sensitivity, or safe behavior. Our experiments are limited to controlled research settings using publicly available multilingual benchmarks, and we do not evaluate deployment in high-risk domains such as legal, medical, or political decision-making. Overall, we view Neural FOXP2 as a research contribution toward controllable multilingual generation and inference-time activation steering rather than a production-ready deployment framework.

References

Lyzander Marciano Andrylie, Inaya Rahmanisa, Mahardika Krisna Ihsani, Alfan Farizki Wicaksono, Haryo Akbarianto Wibowo, and Alham Fikri Aji. 2026. Sparse autoencoders can capture language-specific concepts across diverse languages. *Preprint*, arXiv:2507.11230.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637, Online. Association for Computational Linguistics.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nicholas Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Liane Wu, Shauna Kravec, Nicholas Schiefer, Thomas Maxwell, Nelson Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, and 6 others. 2023. Towards monosemanticity: Decomposing language models with dictionary learning. Transformer Circuits.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee D. Sharkey. 2023. Sparse autoencoders find highly interpretable features in language models. *arXiv preprint arXiv:2309.08600*.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Neel Nanda, Kamal Ndousse, and 7 others. 2022. Toy models of superposition. *arXiv preprint arXiv:2209.10652*.

Yanzhu Guo, Simone Conia, Zelin Zhou, Min Li, Saloni Potdar, and Henry Xiao. 2025. Do large language models have an English accent? evaluating and improving the naturalness of multilingual LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3823–3838, Vienna, Austria. Association for Computational Linguistics.

Songbo Hu, Ivan Vulić, and Anna Korhonen. 2025. Quantifying language disparities in multilingual large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 4003–4018, Suzhou, China. Association for Computational Linguistics.

Takeshi Kojima, Itsuki Okimura, Yusuke Iwasawa, Hitomi Yanaka, and Yutaka Matsuo. 2024a. On the multilingual ability of decoder-based pre-trained language models: Finding and controlling language-specific neurons. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, Mexico City, Mexico. Association for Computational Linguistics.

Takeshi Kojima, Itsuki Okimura, Yusuke Iwasawa, Hitomi Yanaka, and Yutaka Matsuo. 2024b. On the multilingual ability of decoder-based pre-trained language models: Finding and controlling language-specific neurons. *Preprint*, arXiv:2404.02431.

Joanito Agili Lopo, Muhammad Ravi Shulthan Habibi, Tack Hwa Wong, Muhammad Ilham Ghozali, Fajri Koto, Genta Indra Winata, Peerat Limkonchotiwat, Alham Fikri Aji, and Samuel Cahyawijaya. 2025. Language surgery in multilingual large language models. In *Proceedings of the 5th Workshop on Multilingual Representation Learning (MRL 2025)*, pages 438–467, Suzhuo, China. Association for Computational Linguistics.

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

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.

Nina Panickssery, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. 2023. Steering Llama 2 via contrastive activation addition. *arXiv preprint arXiv:2312.06681*.

Inaya Rahmanisa, Lyzander Marciano Andrylie, Mahardika Krisna Ihsani, Alfan Farizki Wicaksono, Haryo Akbarianto Wibowo, and Alham Fikri Aji. 2025. 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*, pages 919–968, Mumbai, India. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.

John Schulman, Sergey Levine, Philipp Moritz, Michael I. Jordan, and Pieter Abbeel. 2017. Trust region policy optimization.

Nadav Shani and Ali Basirat. 2025. Language dominance in multilingual large language models. In *Proceedings of the 8th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*,

pages 137–148, Suzhou, China. Association for Computational Linguistics.

Tianyi Tang, Wenyang Luo, Haoyang Huang, Dongdong Zhang, Xiaolei Wang, Xin Zhao, Furu Wei, and Ji-Rong Wen. 2024. Language-specific neurons: The key to multilingual capabilities in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5701–5715, Bangkok, Thailand. Association for Computational Linguistics.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

Hinata Tezuka and Naoya Inoue. 2025. The transfer neurons hypothesis: An underlying mechanism for language latent space transitions in multilingual LLMs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 31742–31792, Suzhou, China. Association for Computational Linguistics.

Alexander Turner, Lisa Thiergart, Geoffrey Leech, Madeleine Udell, and Chris Olah. 2024. Activation addition: Steering language models without fine-tuning. *arXiv preprint arXiv:2404.04871*.

Xinwei Wu, Heng Liu, Xiaohu Zhao, Yuqi Ren, Linlong Xu, Longyue Wang, Deyi Xiong, Weihua Luo, and Kaifu Zhang. 2026. Finding the translation switch: Discovering and exploiting the task-initiation features in llms. *Preprint*, arXiv:2601.11019.

Fred Zhang and Neel Nanda. 2023. Towards best practices of activation patching in language models: Metrics and methods. *arXiv preprint arXiv:2309.16042*.

Shimao Zhang, Zhejian Lai, Xiang Liu, Shuaijie She, Xiao Liu, Yeyun Gong, Shujian Huang, and Jiajun Chen. 2026. How does alignment enhance llms' multilingual capabilities? a language neurons perspective. *Preprint*, arXiv:2505.21505.

Weixiang Zhao, Jiahe Guo, Yang Deng, Tongtong Wu, Wenxuan Zhang, Yulin Hu, Xingyu Sui, Yanyan Zhao, Wanxiang Che, Bing Qin, Tat-Seng Chua, and Ting Liu. 2025. When less language is more: Language-reasoning disentanglement makes llms better multilingual reasoners. *Preprint*, arXiv:2505.15257.

Chengzhi Zhong, Fei Cheng, Qianying Liu, Yugo Murawaki, Chenhui Chu, and Sadao Kurohashi. 2026. Language lives in sparse dimensions: Toward interpretable and efficient multilingual control for large language models. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7958–7970, Rabat, Morocco. Association for Computational Linguistics.

Appendix

The appendix is structured as follows:

- **(i) Related Work (extended).** We provide a longer, higher-resolution positioning of FOXP2 relative to: (a) **language-specific feature discovery** (e.g., entropy/activation-probability methods), (b) **representation editing** and causal interventions (ROME-style, patching, circuit editing), (c) **multilingual control** via prompting, adapters, and decoding-time methods, and (d) **feature-basis interventions** (e.g., sparse feature dictionaries) that motivate FOXP2s coordinate choice. We emphasize the methodological distinction between *detecting* language selectivity and *constructing* a stable, low-rank control artifact (N, S, W) (N, S, W). cf. appendix A.

- **(ii) Datasets and Experimental Details.** We document the complete experimental harness: model checkpoints and tokenizer hashes; prompt sources and splits (neutral prompts, plus any safety/refusal slice); decoding settings; batching; hardware; and all randomness controls. We also specify the matched meaning-unit construction used to compute ΔZ (pairing logic, filters, and how we avoid spurious topical confounds). cf. appendix C.

- **(iii) FOXP2: Full Algorithmic Specification.** We present FOXP2 as a three-stage pipeline that outputs (N, S, W) and an edit rule, with precise algorithms, hyperparameters, and pass/fail sanity checks:

  - **(iii-a) Step I: Localize (discover N_{ℓ_t}).** We define the candidate coordinate space, the target selectivity / causal-lift scoring used to rank features, and the deterministic thresholding procedure that yields a compact support $\mathcal{N}_{\ell_t}$. We include stability diagnostics (bootstrap agreement, seed sensitivity) and the explicit necessity ablation that removes $\mathcal{N}_{\ell_t}$ during editing. cf. appendix B.
  - **(iii-b) Step II: Discover low-rank directions and window (construct $S_{\ell_t}^{(\ell)}$ and W).** We specify how $\Delta Z^{(\mathcal{W},\ell_t)}$ is formed from matched meaning units, how SVD is computed, and how we report and use **effective rank** r_{eff}, **eigengap** σ_r/σ_{r+1}, and **bootstrap principal-angle stability**. We then define the window-selection objective and provide the algorithm that selects the final intervention window W and per-layer subspaces $S_{\ell_t}^{(\ell)}$. cf. appendix B.
  - **(iii-c) Step III: Steer (signed sparse edit).** We give the exact edit equations and hook placement: how the intervention is restricted to $\mathcal{N}_{\ell_t}$, how it projects onto $S_{\ell_t}^{(\ell)}$, and how it implements **suppression of dominant English-biased directions** without degenerating into generic logit flattening. We also specify entropy-matched controls (temperature / isotropic noise) and sufficiency tests (edit only on $\mathcal{N}_{\ell_t}$ reproduces gains). cf. appendix B.

- **(iv) Metrics: Definitions, Validators, and Robustness Checks.** We provide complete metric specifications: token-mass defaultness $M^{\ell}(x, t)$, ΔM, horizon sweep protocol (and why early horizons operationalize defaultness), token-set construction variants, shared-token handling, and the cross-metric validation channel (script-agnostic LID + human anchoring). We also define ΔS (utility) as a matched-suite delta with bootstrap confidence intervals, and we detail factuality/calibration and safety invariance checks. cf. appendix D.

- **(v) Results: Extended Tables, Curves, and Ablations.** We report additional results beyond the main text: horizon sweep stability (without hiding longer horizons), token-set sensitivity bands, dropout perturbations on V_{ℓ}, necessity/sufficiency ablations, entropy-matched control comparisons, domain-transfer retention curves, and system-level overhead (throughput/latency/memory) relative to prompt-only and small LoRA. We include additional error analyses and qualitative failure modes where defaultness gains do not translate to human natural start. cf. appendix E.

This Appendix is designed to serve as the **reproducibility and audit reference** for FOXP2: any reader should be able to reconstruct the (N, S, W) artifact, rerun the intervention-based evaluations, and verify defaultness gains and utility/safety invariance under the complete evaluation harness.

A Related Work: Language-Specific Features and Multilingual Steering (Extended)

A growing line of work argues that multilingual capability in LLMs is mediated by **language-selective internal circuitry** and that **intervening** on this circuitry can steer the output language. Tang et al. (Tang et al., 2024) operationalize this view by proposing **LAPE** (*Language Activation Probability Entropy*) to detect *language-specific features* in FFN blocks, and by showing that manually activating or deactivating such features can steer output language. FOXP2 is aligned with the mechanistic spirit of this agenda, but differs **mathematically and algorithmically** in *what* it estimates, *which coordinates* it treats as causal, *how* it constructs steering directions, and *what* it certifies under deployment constraints.

Detection objective versus control objective. Tang et al. seek units that are **selective to a language** in the sense of activation probability across corpora. Let $a_j^{(\ell)}(x) \in \mathbb{R}$ denote the post-nonlinearity activation of FFN feature j at layer ℓ for input x. Define an activation event, for example $\mathbb{I}[a_j^{(\ell)}(x) > 0]$. Over a corpus $\mathcal{D}_k$ in language k, estimate

$$p_{j,k}^{(\ell)} \triangleq \Pr_{x \sim \mathcal{D}_k} \big[a_j^{(\ell)}(x) > 0\big],$$

$$q_{j,k}^{(\ell)} \triangleq \frac{p_{j,k}^{(\ell)}}{\sum_{k'} p_{j,k'}^{(\ell)}},$$

$$H_j^{(\ell)} \triangleq -\sum_k q_{j,k}^{(\ell)} \log q_{j,k}^{(\ell)}.$$

Low $H_j^{(\ell)}$ indicates language-specific features, which can then be toggled to steer outputs (Tang et al., 2024). This is a distributional selectivity criterion: it identifies where language-skewed units reside and whether their activation patterns correlate with language.

FOXP2 targets a different estimand: **language defaultness**, defined as the models **initial commitment** to a language before content constraints dominate. We operationalize defaultness through an early-horizon logit-mass event. Let $p_\theta(\cdot \mid \mathrm{ctx}_t)$ be the next-token distribution at step t, and let V_ℓ be the token set for language ℓ. Define

$$M^\ell(x,t) \triangleq \sum_{u \in V_\ell} p_\theta\big(u \mid \mathrm{ctx}_t(x)\big),$$

$$\Delta M(x,t) \triangleq M^{\ell_t}(x,t) - M^{\mathrm{en}}(x,t).$$

FOXP2 reports an early commitment score aggregated over $t \in \{1,2,3\}$. Thus, FOXP2 does not ask only which features(derived from neurons) are language-specific under corpus statistics; it asks whether probability mass can be causally and stably reallocated so that the model begins in the target language by default, with bounded collateral damage.

Coordinate choice. Tang et al. intervene on **raw FFN neurons** by changing activations toward language-conditioned averages. If $a^{(\ell)}(x) \in \mathbb{R}^{d_{\mathrm{ff}}}$ is the FFN post-activation vector, their intervention can be written as

$$a^{(\ell)}(x) \leftarrow a^{(\ell)}(x) + \eta \cdot \mathbf{m}_k^{(\ell)} \odot \big(\bar{a}_k^{(\ell)} - a^{(\ell)}(x)\big),$$

where $\mathbf{m}_k^{(\ell)} \in \{0,1\}^{d_{\mathrm{ff}}}$ masks detected language-specific features for language k, and $\bar{a}_k^{(\ell)}$ is a language-specific mean activation vector estimated from corpora (Tang et al., 2024). This is direct neuron-level actuation.

FOXP2 instead edits in a **sparse feature basis**. Let $h_t^{(\ell)} \in \mathbb{R}^d$ be the residual-stream representation at layer ℓ, and let $D^{(\ell)} \in \mathbb{R}^{d \times m}$ be a sparse autoencoder dictionary with encoder $E^{(\ell)}$, so that

$$z_t^{(\ell)} = E^{(\ell)} h_t^{(\ell)}, h_t^{(\ell)} \approx D^{(\ell)} z_t^{(\ell)},$$

with sparse $z_t^{(\ell)}$. FOXP2 identifies a support $\mathcal{N}_{\ell_t} \subseteq [m]$ of language-relevant coordinates and intervenes only on those coordinates, mapping back through $D^{(\ell)}$. This matters under superposition: the same native neuron may participate in multiple features, while sparse features provide a cleaner handle for localized edits. Tang et al. actuate native units; FOXP2 actuates identified sparse features.

Prompt-only methods, dense interventions, and the FOXP2 control object. FOXP2 differs from prompt-only multilingual interventions because prompting operates only at the input level and cannot modify internal representations. Our results indicate that language defaultness is embedded in the models residual activations, so input-level instructions alone cannot fully remove this bias.

FOXP2 also differs from dense cross-lingual intervention methods that apply a broad shift over activations. Instead, FOXP2 computes steering directions within a localized feature support. Given feature vectors $z^{(\ell)}(x)$ extracted at a designated

token position, we form a displacement matrix

$$\Delta Z^{(\ell)} \triangleq \begin{bmatrix} z^{(\ell)}(x_1^{\ell_t}) - z^{(\ell)}(x_1^{\mathrm{en}}) \\ \vdots \\ z^{(\ell)}(x_n^{\ell_t}) - z^{(\ell)}(x_n^{\mathrm{en}}) \end{bmatrix} \in \mathbb{R}^{n \times m},$$

and compute an SVD

$$\Delta Z^{(\ell)} = U^{(\ell)} \Sigma^{(\ell)} (V^{(\ell)})^\top.$$

The candidate steering subspace is

$$S_{\ell_t}^{(\ell)}(r) \triangleq \mathrm{span}\{v_1^{(\ell)}, \ldots, v_r^{(\ell)}\},$$

where $v_i^{(\ell)}$ are the right singular vectors. We choose r using eigengap and bootstrap stability, rather than treating the shift as a dense global edit. The low-rank-only ablation underperforms and shows higher leakage, while the full method achieves low leakage, indicating that the localized feature support is important for precision and reduced side effects.

Intervention form. Tang et al. use additive activation steering: once language-specific features are identified, they are turned up or down to encourage the model to output the desired language (Tang et al., 2024). FOXP2 instead uses a signed, sparse, subspace-aligned edit designed to implement directed probability-mass reallocation at the earliest steps:

$$\begin{aligned} z^{(\ell)} \leftarrow z^{(\ell)} &+ \lambda\, P_{S_{\ell_t}^{(\ell)}}\big(\Pi_{\mathcal{N}_{\ell_t}} z^{(\ell)}\big) \\ &- \lambda'\, P_{S_{\mathrm{en}}^{(\ell)}}\big(\Pi_{\mathcal{N}_{\mathrm{en}}} z^{(\ell)}\big), \qquad \ell \in W. \end{aligned}$$

where $\Pi_{\mathcal{N}}$ projects onto the sparse support and P_S projects onto the steering subspace. This bipolar control law explicitly promotes the target language while suppressing the English-default attractor. We validate this behavior with entropy-matched controls such as temperature scaling or isotropic noise.

Relation to prompt-only, dense, and suppression-based multilingual interventions. Recent training-free multilingual intervention methods should not be treated as a uniform class because they differ fundamentally in their *intervention object*, *control objective*, and *assumptions about multilingual representations* (Zhao et al., 2025). FOXP2 is best understood relative to these axes. Prompt-only methods operate entirely at the input level, relying on natural-language instructions or demonstrations to bias the model toward a target language without modifying internal computation. Dense activation steering methods instead intervene directly in hidden representations, typically by adding global language vectors or residual directions computed from bilingual activation differences. More recent suppression-oriented approaches argue that multilingual reasoning quality can improve when language identity signals are partially disentangled or attenuated during inference (Zhao et al., 2025). FOXP2 differs from all three categories by treating multilingual behavior as arising from sparse, selectively activated neuron subspaces rather than dense distributed language directions alone. Instead of prompting the model externally or globally shifting all residual activations, FOXP2 identifies language-selective sparse autoencoder features and performs localized causal steering over neuron subsets associated with the target language. This yields a mechanistic intervention framework that is sparse, interpretable, and layer-specific while remaining training-free at deployment time.

FOXP2 differs from prompt-only multilingual steering methods because prompting operates purely at the input level and does not modify internal representations. Prompting can bias the decoding trajectory through instruction conditioning, but it cannot directly reshape the latent representational geometry responsible for language defaultness. Our experiments indicate that defaultness is encoded in residual-stream activations and manifests as an early probability-mass bias toward English. As a result, instruction prompting alone cannot reliably overcome this attractor, especially in early decoding horizons where the model commits to a language before semantic constraints dominate generation.

FOXP2 also differs from dense cross-lingual intervention methods that apply broad activation shifts over hidden states or residual streams. Such approaches typically assume that multilingual behavior can be modified through a global representational displacement. In contrast, FOXP2 first localizes a sparse support of language-relevant features and only then computes low-rank steering directions within that support. The intervention therefore acts on a geometrically constrained subspace rather than on the entire representation manifold. Empirically, this distinction matters: the low-rank-only ablation underperforms and exhibits substantially higher leakage, whereas the full

sparse-plus-low-rank formulation produces significantly cleaner control with reduced cross-language spillover. This suggests that localization is necessary for precision and that low-rank structure alone is insufficient.

FOXP2 is also conceptually distinct from *When Less Language Is More*, which studies whether suppressing language-specific representations can improve multilingual reasoning performance. That line of work treats language-specific structure primarily as a nuisance variable whose removal may improve abstraction or transfer. FOXP2 addresses a different problem. Rather than suppressing language realization, FOXP2 seeks to strengthen a target-language realization channel while simultaneously suppressing the English-default attractor. Concretely, the intervention is signed:

$$z^{(\ell)} \leftarrow z^{(\ell)} + \lambda\, P_{S_{\ell_t}^{(\ell)}}\big(\Pi_{\mathcal{N}_{\ell_t}} z^{(\ell)}\big) - \lambda'\, P_{S_{\mathrm{en}}^{(\ell)}}\big(\Pi_{\mathcal{N}_{\mathrm{en}}} z^{(\ell)}\big).$$

so the objective is not merely to remove language information, but to causally redirect early decoding dynamics toward a desired language manifold while preserving semantic utility. Additionally, FOXP2 introduces an explicit **defaultness metric** based on early-horizon probability-mass transfer ΔM, whereas prior work typically evaluates language effects indirectly through downstream task accuracy or emergent multilingual behavior.

Finally, FOXP2 is distinct in combining three ingredients simultaneously:

1. SAE-based sparse feature localization,
2. low-rank steering restricted to that localized support, and
3. defaultness-specific evaluation with causal controls, including random-feature edits, ablations, and leakage analysis.

These controls show that the observed gains arise from a localized and structured intervention pathway rather than from generic activation perturbation.

What the ablations show. Compared to LAPE-style steering, FOXP2 avoids operating on raw activation units, which are often polysemantic. The random-feature ablation in Table 1 shows that arbitrary feature edits produce negligible gains, while the full method achieves a strong improvement, showing that specific localized features, not generic activation edits, drive the effect. Compared to the low-rank-only variant, the full method performs better and leaks less, showing that low-rank directions are effective only when computed within the localized support. Taken together, these controls support the view that the observed gains arise from a structured causal pathway rather than from broad activation steering.

What is claimed and what is certified. Tang et al. provide mechanistic evidence that language-specific features exist, are concentrated in certain layers, and can be toggled to steer output language (Tang et al., 2024). FOXP2 goes further and aims to provide a reproducible control artifact with explicit validity and deployment constraints. The core contribution is not merely that language-selective units exist; it is that defaultness can be controlled reproducibly by extracting a stable low-rank actuator and executing a signed sparse edit under explicit guardrails.

A.1 Baseline Instantiation: LAPE Language-Specific Neuron Steering under the FOXP2 Evaluation Harness

A natural reviewer question is whether FOXP2s gains arise simply because language-specific units exist and can be toggled (Tang et al., 2024). To answer this, we instantiate a faithful LAPE-based steering baseline inside our measurement regime: language defaultness (early-horizon mass shift plus script-agnostic LID), leakage, subspace stability, and utility. This appendix specifies the conversion precisely so that the comparison is auditable and does not conflate evaluation definitions.

LAPE detection. Let $a_j^{(\ell)}(x) \in \mathbb{R}$ denote the post-activation value of FFN neuron j at layer ℓ for input x, and define an active event $\mathbb{I}[a_j^{(\ell)}(x) > 0]$. Given corpora $\{\mathcal{D}_k\}_{k=1}^{K}$ for languages k, estimate

$$p_{j,k}^{(\ell)} \triangleq \Pr_{x\sim\mathcal{D}_k}\Big[a_j^{(\ell)}(x) > 0\Big],\; q_{j,k}^{(\ell)} \triangleq \frac{p_{j,k}^{(\ell)}}{\sum_{k'} p_{j,k'}^{(\ell)}}.$$

Then compute Language Activation Probability Entropy

$$H_j^{(\ell)} \triangleq -\sum_{k=1}^{K} q_{j,k}^{(\ell)} \log q_{j,k}^{(\ell)}.$$

Low $H_j^{(\ell)}$ identifies a features whose activation probability is concentrated in a subset of languages.

Table 3: Comparative summary: Tang et al. (ACL'24) vs. FOXP2 (this work).

Axis	Tang et al. (ACL'24)	FOXP2 (this work)
Primary question	Which FFN neurons are language-specific?	Can we control defaultness at inference-time?
Estimand	Activation-probability selectivity across languages	Early-horizon mass transfer ΔM with LID validation
Coordinates	Native FFN neurons	SAE feature coordinates (superposition-aware)
Geometry	Per-unit entropy; layer localization	SVD subspaces $S_{\ell_t}^{(\ell)}$, eigengaps, principal angles
Intervention	Activate or deactivate neuron sets	Signed sparse edit: promote target language, suppress English in low-rank subspaces
Stability criterion	Language skew in corpora	Bootstrap subspace stability and window selection
Control artifact	Detected neuron lists as analysis outcome	Reproducible $(\mathcal{N}, \mathcal{S}, W)$ and deployment contract
Deployment stance	Feasibility demonstration	KV-cache, batching, spec-dec constraints, and overhead reporting

For each target language $\ell_t \in \{\mathrm{hi}, \mathrm{es}\}$, we construct a mask

$$m_{\ell_t}^{(\ell)}(j) = 1 \iff H_j^{(\ell)} \le \tau_H \wedge \ell_t = \arg\max_k q_{j,k}^{(\ell)}.$$

From detection to steering. We implement the baseline as a masked additive shift at the same token position and within the same layer window used by FOXP2:

$$a_t^{(\ell)}(x) \leftarrow a_t^{(\ell)}(x) + \eta_{\ell_t}\big(m_{\ell_t}^{(\ell)} \odot s_{\ell_t}^{(\ell)}\big), \qquad \ell \in W.$$

Two faithful choices are standard: mean-target shift

$$s_{\ell_t}^{(\ell)} \triangleq \mu_{\ell_t}^{(\ell)} - \mu_{\mathrm{en}}^{(\ell)},$$

or a unit push

$$s_{\ell_t}^{(\ell)} \equiv \mathbf{1}$$

restricted to the mask. We report the exact variant used so that the baseline is fully specified.

Budget matching and scoring. To avoid penalizing or advantaging the baseline, we match intervention budget in two ways: count-matched and FLOP/memory-matched. The baseline is then scored under the same metrics as FOXP2: early-horizon token mass, script-agnostic language ID, leakage, and utility. This lets us test whether FOXP2s gains come from the structured sparse-feature plus low-rank control design, rather than from generic activation steering.

Why the comparison is informative. LAPE identifies selective units through activation statistics, but it does not enforce low-rank actuator geometry, bootstrap stability across prompt families, or bipolar mass transfer. As a result, under our harness it may show weaker early defaultness gains, higher leakage, or larger utility loss for a similar intervention budget. FOXP2 is designed specifically to address these issues through a reproducible control artifact and explicit guardrails.

B FOXP2: Full Algorithmic Specification

This section specifies **Neural FOXP2** as a complete three-stage pipeline that discovers $(\mathcal{N}_{\ell_t}, \mathcal{S}, \mathcal{W})$ and yields a deterministic inference-time edit rule with explicit **guardrails**, **stability gates**, and **matched-controls** that rule out degenerate "generic logit reshaping." FOXP2 targets **defaultness as early commitment**: it aims to shift the model's *initial* language choice (first T tokens) while keeping semantics and non-target language behavior stable (Appendix §D). The full pipeline is:

$$\boxed{\begin{aligned}&\textbf{FOXP2}: \ \big(\theta, \{\mathrm{SAE}^{(\ell)}\}_{\ell=1}^{L}, \mathcal{D}_{\mathrm{disc}}, \ell_t\big)\\ &\longrightarrow \ \big(\mathcal{N}_{\ell_t}, \{\mathcal{S}_{\ell_t}^{(\ell)}\}_{\ell\in\mathcal{W}}, \mathcal{W}, \mathtt{EditRule}_{\ell_t}\big)\end{aligned}}$$

We present: **(I) Localize** (discover sparse coordinates $\mathcal{N}_{\ell_t}$), **(II) Discover** (low-rank subspace $\mathcal{S}_{\ell_t}^{(\ell)}$ and intervention window $\mathcal{W}$), and **(III) Steer** (signed sparse edit with drift-regularized suppression of dominant English-biased directions), plus explicit **sanity checks** and **pass/fail criteria**.

Preliminaries: what FOXP2 edits, and why SAE feature space

Model states and decoding steps. Fix a decoder-only transformer with L layers. Given a prompt x, decoding step t produces a prefix $y_{<t}$ and an internal residual-stream hidden state $h_\theta^{(\ell)}(x,t) \in \mathbb{R}^d$ at layer ℓ. The next-token distribution is $p_\theta(\cdot \mid x, y_{<t})$.

SAE feature coordinates. For each layer ℓ we assume a frozen SAE with encoder $\mathrm{Enc}^{(\ell)} : \mathbb{R}^d \to \mathbb{R}^m$ and decoder $\mathrm{Dec}^{(\ell)} : \mathbb{R}^m \to \mathbb{R}^d$ (fixed after pretraining on θ's activations). Feature activations are:

$$\boxed{z_\theta^{(\ell)}(x,t) = \mathrm{Enc}^{(\ell)}\big(h_\theta^{(\ell)}(x,t)\big) \in \mathbb{R}^m.}$$

FOXP2 edits are applied in z-space and mapped back via $\mathrm{Dec}^{(\ell)}$:

$$\boxed{\begin{aligned}\Delta h^{(\ell)}(x,t) &= \mathrm{Dec}^{(\ell)}\big(\Delta z^{(\ell)}(x,t)\big),\\ h_{\theta'}^{(\ell)}(x,t) &= h_\theta^{(\ell)}(x,t) + \Delta h^{(\ell)}(x,t).\end{aligned}}$$

This induces structured edits that are (i) coordinate-localizable (sparse supports), (ii) low-rank-compressible (subspaces), and (iii) stable to resampling (bootstrap principal-angle tests).

Meaning units (paired semantic contexts). FOXP2 requires matched semantic content so that discovery isolates a language-control mechanism rather than topic/format effects. A **meaning unit** is a prompt family where task content is fixed but the *language condition* differs minimally. For each semantic item x, we create two conditions: $\mathrm{cond} = \ell_t$ (target-language cue) and $\mathrm{cond} = \mathrm{en}$ (English cue), holding the remainder identical. We denote paired runs by (x, ℓ_t) and (x, en). Extraction is done for a short horizon T_0 (default $T_0 \in \{1,2,3\}$), because defaultness is an early commitment phenomenon.

Discovery vs evaluation splits. $\mathcal{D}_{\mathrm{disc}}$ is used only to discover $(\mathcal{N}, \mathcal{S}, \mathcal{W})$ and tune the operating point $\lambda^\star$; all reporting uses held-out $\mathcal{D}_{\mathrm{eval}}$ (Appendix §C). Bootstrap for CIs resamples prompts, not tokens.

Two-channel construct validity (hard requirement). FOXP2 claims are credited only when (i) the mechanistic token-mass channel and (ii) a script-agnostic LID validator agree in direction under fixed decoding (Appendix §D). Mass-only improvements with flat/negative LID are treated as script/token-set artifacts.

Step I: Localize (discover sparse support $\mathcal{N}_{\ell_t}$)

Step I finds a compact set of SAE coordinates that are (a) **target-selective** under matched meaning units and (b) **causally useful** for early-horizon defaultness. The output is a deterministic sparse support $\mathcal{N}_{\ell_t} \subseteq \{(\ell, i) : \ell \in [L], i \in [m]\}$.

Candidate coordinate space and paired differences

Fix candidate layers $\mathcal{L}_{\mathrm{cand}}$ (wide by default) and horizon T_0. For each $x \in \mathcal{D}_{\mathrm{disc}}$ and $t \le T_0$, compute SAE activations under both conditions:

$$z_\theta^{(\ell)}(x, t \mid \ell_t), \qquad z_\theta^{(\ell)}(x, t \mid \mathrm{en}).$$

Define paired feature differences:

$$\boxed{\begin{aligned}&\Delta z^{(\ell)}(x,t) = z_\theta^{(\ell)}(x, t \mid \ell_t) - z_\theta^{(\ell)}(x, t \mid \mathrm{en}),\\ &\Delta z^{(\ell)}(x,t) \in \mathbb{R}^m.\end{aligned}}$$

For feature i, denote $\Delta z_i^{(\ell)}(x,t)$.

Target-selectivity scoring (correlational, but reproducible)

We aggregate paired differences across prompts and early steps:

$$\mu_i^{(\ell)} = \frac{1}{|\mathcal{D}_{\mathrm{disc}}| T_0} \sum_{x \in \mathcal{D}_{\mathrm{disc}}} \sum_{t=1}^{T_0} \Delta z_i^{(\ell)}(x,t),$$

$$\nu_i^{(\ell)} = \sqrt{\frac{1}{|\mathcal{D}_{\mathrm{disc}}| T_0} \sum_{x,t} \big(\Delta z_i^{(\ell)}(x,t) - \mu_i^{(\ell)}\big)^2} + \epsilon.$$

Define a deterministic selectivity statistic (a signal-to-noise ratio in paired-difference space):

$$\boxed{\mathrm{Sel}_i^{(\ell)} = \frac{\mu_i^{(\ell)}}{\nu_i^{(\ell)}}.}$$

Interpretation: $\mathrm{Sel}_i^{(\ell)}$ is large if feature i shifts consistently positive when swapping English-cue $\to$ target-cue under matched content.

Anti-confound filtering (shared/format sensitivity). FOXP2 is designed to avoid features that are primarily sensitive to punctuation, digits, JSON templates, or shared tokens. We thus compute (on a small probe set) a **format sensitivity** score by repeating the paired protocol across multiple format wrappers (plain, bullets, JSON) while keeping meaning fixed; features with high variance across formats relative to $\mu_i^{(\ell)}$ are flagged and optionally excluded from candidates before top-k selection. This is a deterministic pre-filter (no extra learning).

Causal-lift scoring (do-proxy via micro-interventions)

Selectivity can capture correlates; FOXP2 additionally tests causal usefulness by micro-intervention on individual features. Let $M^{\ell_t}_{\theta,T_0}(x)$ be the weighted token-mass metric for the target language at horizon T_0 (Appendix §D). For each candidate coordinate (ℓ, i), define a micro-edit:

$$z_i^{(\ell)}(x,t) \leftarrow z_i^{(\ell)}(x,t) + \delta$$

applied at the hook site and only for $\ell \in \mathcal{L}_{\text{cand}}$, $t \le T_0$. The causal lift is:

$$\boxed{\text{Lift}_i^{(\ell)} = \frac{1}{|\mathcal{D}_{\text{cal}}|} \sum_{x \in \mathcal{D}_{\text{cal}}} \Big(M^{\ell_t}_{\theta+\text{do}(\ell,i,\delta),T_0}(x) - M^{\ell_t}_{\theta,T_0}(x) \Big),}$$

where $\mathcal{D}_{\text{cal}} \subset \mathcal{D}_{\text{disc}}$ is a small calibration subset. This is a *do-proxy*: we explicitly intervene on a coordinate and test the induced change in the mechanistic defaultness probe.

Why lift is measured on token-mass, not LID. Lift is computed on the mass channel because LID is detector-based and less sensitive at token-1–3. Construct validity is enforced later at the method level, not at single-feature lift level.

Composite ranking and deterministic selection of $\mathcal{N}_{\ell_t}$

We combine selectivity and lift into a composite score:

$$\text{Score}_i^{(\ell)} = \text{Sel}_i^{(\ell)} + \alpha_{\text{lift}} \cdot \frac{\text{Lift}_i^{(\ell)}}{\widehat{\sigma}_{\text{Lift}} + \epsilon},$$

where $\widehat{\sigma}_{\text{Lift}}$ is the empirical standard deviation of lift over candidates. The sparse support is chosen by top-k selection:

$$\boxed{\mathcal{N}_{\ell_t} = \text{TopK}\Big(\{(\ell,i) : \ell \in \mathcal{L}_{\text{cand}}, i \in [m]\}, \text{Score}, k\Big).}$$

We optionally enforce per-layer caps ($\le k_{\max}$ per layer) to prevent single-layer collapse and to keep the support distributed.

Bootstrap reproducibility and sign-stability gates

For bootstrap $b = 1, \ldots, B$, resample $\mathcal{D}_{\text{disc}}$ with replacement and recompute $\mathcal{N}^{(b)}_{\ell_t}$. We report overlap stability:

$$\boxed{\text{Stab}_{\mathcal{N}} = \frac{1}{B} \sum_{b=1}^{B} \frac{|\mathcal{N}^{(b)}_{\ell_t} \cap \mathcal{N}^{(0)}_{\ell_t}|}{|\mathcal{N}^{(0)}_{\ell_t}|}.}$$

Additionally, for each selected coordinate (ℓ, i) we compute **sign stability** $\Pr_b[\mu_i^{(\ell,b)} > 0]$. Coordinates with sign stability below a fixed threshold (e.g., < 0.9) are deterministically removed and replaced by the next-highest scoring candidates, preserving $|\mathcal{N}_{\ell_t}| = k$.

Necessity ablation (explicit, deployment-relevant)

We run the w/o $\mathcal{N}_{\ell_t}$ ablation by applying the full Step III edit rule but forcing $\Pi_{\mathcal{N}_{\ell_t}} = 0$ (equivalently, $\mathcal{N}_{\ell_t} = \varnothing$). The resulting losses in DEFAULTHI, $\Delta^{\text{hi}}_{\text{mass}}$, and $\Delta^{\text{hi}}_{\text{lid}}$ quantify necessity (Table 9).

Step II: Discover low-rank directions and window (construct $\mathcal{S}^{(\ell)}_{\ell_t}$ and $\mathcal{W}$)

Step II compresses the localized sparse mechanism into a low-rank structure and localizes where it is *causally effective*. It outputs (i) per-layer target subspaces $\{\mathcal{S}^{(\ell)}_{\ell_t}\}$ and (ii) a contiguous intervention window $\mathcal{W}$.

Restrict paired differences to the localized support

Define a coordinate projection $\Pi_{\mathcal{N}_{\ell_t}}$ in feature space (zeros outside $\mathcal{N}_{\ell_t}$). For each layer ℓ and each paired sample (x, t), define:

$$\boxed{\Delta\tilde{z}^{(\ell)}(x,t) = \Pi_{\mathcal{N}_{\ell_t}} \Delta z^{(\ell)}(x,t) \in \mathbb{R}^m.}$$

Stack these vectors to form a difference matrix:

$$\boxed{\Delta Z^{(\ell)} = \begin{bmatrix} (\Delta\tilde{z}^{(\ell)}(x_1,1))^\top \\ \vdots \\ (\Delta\tilde{z}^{(\ell)}(x_n,T_0))^\top \end{bmatrix} \in \mathbb{R}^{(nT_0)\times m}.}$$

Interpretation: rows are matched meaning-unit displacements in feature space, supported only on localized coordinates.

SVD, dominant directions, and a per-layer steering subspace

Compute SVD:

$$\Delta Z^{(\ell)} = U^{(\ell)}\Sigma^{(\ell)}(V^{(\ell)})^\top,$$
$$\Sigma^{(\ell)} = \mathrm{diag}(\sigma_1^{(\ell)} \geq \sigma_2^{(\ell)} \geq \cdots).$$

Define the rank-r subspace:

$$\boxed{\mathcal{S}_{\ell_t}^{(\ell)} = \mathrm{span}\big(V_{:,1:r}^{(\ell)}\big) \subset \mathbb{R}^m.}$$

This is the low-rank structure FOXP2 uses: edits will be projected onto $\mathcal{S}_{\ell_t}^{(\ell)}$.

Effective rank and eigengap (quantifying "sparse + low-rank")

Let $p_j^{(\ell)} = \sigma_j^{(\ell)} / \sum_k \sigma_k^{(\ell)}$. The effective rank is:

$$\boxed{r_{\mathrm{eff}}^{(\ell)} = \exp\Big(-\sum_j p_j^{(\ell)} \log p_j^{(\ell)}\Big).}$$

We also report an eigengap ratio at r:

$$\boxed{\mathrm{Gap}_r^{(\ell)} = \frac{\sigma_r^{(\ell)}}{\sigma_{r+1}^{(\ell)} + \epsilon}.}$$

A FOXP2-consistent mechanism exhibits small $r_{\mathrm{eff}}^{(\ell)}$ and a large $\mathrm{Gap}_r^{(\ell)}$ in the effective window.

Bootstrap principal-angle stability of the subspace

For each bootstrap b, recompute $\Delta Z^{(\ell,b)}$ and its top-r subspace $\mathcal{S}_{\ell_t}^{(\ell,b)}$. Measure agreement with a reference $\mathcal{S}_{\ell_t}^{(\ell,0)}$ via principal angles $\angle_i$:

$$\boxed{\begin{aligned}\mathrm{Stab}_{\mathcal{S}}^{(\ell)} = 1 - \frac{1}{B}\sum_{b=1}^{B}\frac{1}{r} \\ \times \sum_{i=1}^{r} \sin\big(\angle_i(\mathcal{S}_{\ell_t}^{(\ell,b)}, \mathcal{S}_{\ell_t}^{(\ell,0)})\big).\end{aligned}}$$

Layers with low $\mathrm{Stab}_{\mathcal{S}}^{(\ell)}$ are not eligible for $\mathcal{W}$.

Window selection (where edits are causally effective)

FOXP2 selects a contiguous window $\mathcal{W}$ by maximizing a dev objective that balances gain, leakage, drift, and utility. For a candidate window $W = [a, b]$, define the dev objective:

$$\boxed{\begin{aligned}\mathcal{J}(W) = \mathrm{Gain}^{\mathrm{hi}}(W) \\ - \alpha_{\mathrm{leak}}\mathrm{Leak}_{\mathrm{hi}\to\mathrm{es}}(W) \\ - \alpha_{\mathrm{kl}}\mathbb{E}[\mathrm{KL}_{T_0}](W) \\ - \alpha_{\mathrm{util}}|\Delta S(W)| \\ - \alpha_{\mathrm{stab}}\sum_{\ell\in W}(1-\mathrm{Stab}_{\mathcal{S}}^{(\ell)}).\end{aligned}}$$

Here $\mathrm{Gain}^{\mathrm{hi}}(W)$ can be computed using DEFAULTHI or the continuous mass channel; leakage/drift/utility are exactly those used in the appendix tables. We search over a candidate set $\mathcal{C}$ of windows (width constraints and centers) and select:

$$\boxed{\begin{aligned}\mathcal{W} = \arg\max_{W\in\mathcal{C}} \mathcal{J}(W) \\ \text{s.t. guardrails pass on dev.}\end{aligned}}$$

This explicitly separates (i) *signal localization* (Step I/II) from (ii) *causal edit effectiveness* (Step III measured during window search). Figure 12 reports the corresponding sweep.

Step III: Steer (signed sparse edit with drift-regularized negative English suppression)

Step III defines the actual inference-time intervention. It is **sparse** (restricted to $\mathcal{N}_{\ell_t}$), **low-rank** (projected to $\mathcal{S}_{\ell_t}^{(\ell)}$), **windowed** (applied only for $\ell \in \mathcal{W}$), and **drift-controlled** (KL trust region + matched controls). We give the exact edit equations, the choice of the English suppression coefficient, and the operational point selection.

Projectors in feature space

Let $V_r^{(\ell)} = V_{:,1:r}^{(\ell)} \in \mathbb{R}^{m\times r}$. The orthogonal projector onto $\mathcal{S}_{\ell_t}^{(\ell)}$ is $P_{\mathcal{S}}^{(\ell)} = V_r^{(\ell)}(V_r^{(\ell)})^\top$. Define the sparse mask $\Pi_{\mathcal{N}_{\ell_t}}$ (diagonal 0/1 in coordinate space). The composed projector is:

$$\boxed{\mathcal{P}_{\ell_t}^{(\ell)} = \Pi_{\mathcal{N}_{\ell_t}} P_{\mathcal{S}}^{(\ell)} \Pi_{\mathcal{N}_{\ell_t}}.}$$

Edits are applied only through $\mathcal{P}_{\ell_t}^{(\ell)}$, enforcing the "sparse + low-rank" structure in the actual intervention.

Data-derived target direction (per layer)

Define a mean displacement direction in the discovered subspace:

$$v_{\ell_t}^{(\ell)} = \mathrm{Norm}\Bigg(\frac{1}{|\mathcal{D}_{\mathrm{disc}}|T_0} \times \sum_{x,t} \mathcal{P}_{\ell_t}^{(\ell)}\big(z_\theta^{(\ell)}(x,t \mid \ell_t) - z_\theta^{(\ell)}(x,t \mid \mathrm{en})\big)\Bigg),$$

where $\mathrm{Norm}(u) = u/(\|u\|_2 + \epsilon)$. This ensures λ is comparable across layers.

Negative English suppression as a drift-regularized compensation (not generic flattening)

A naive "push target" edit can inadvertently increase entropy or broadly reshape the distribution, creating spurious token-mass gains or inflating non-target languages. FOXP2 therefore includes an explicit, structured suppression term that is (i) in the same coordinate support and (ii) in the same low-rank subspace, so it does not become a global logit smoothing trick. Define the corresponding English-associated direction:

$$v_{\mathrm{en}}^{(\ell)} = \mathrm{Norm}\Bigg(\frac{1}{|\mathcal{D}_{\mathrm{disc}}|T_0} \times \sum_{x,t} \mathcal{P}_{\ell_t}^{(\ell)}\big(z_\theta^{(\ell)}(x,t \mid \mathrm{en}) - z_\theta^{(\ell)}(x,t \mid \ell_t)\big)\Bigg).$$

The per-layer feature edit is:

$$\Delta z_{\mathrm{FOXP2}}^{(\ell)}(x,t;\lambda) = \lambda\, v_{\ell_t}^{(\ell)} - \kappa(\lambda)\, v_{\mathrm{en}}^{(\ell)}, \quad \ell \in \mathcal{W}.$$

The key is how $\kappa(\lambda)$ is chosen: FOXP2 chooses $\kappa(\lambda)$ to **minimize drift** subject to achieving gain, rather than letting suppression freely distort logits.

Constrained choice of $\kappa(\lambda)$ (first-order view). Let $\mathrm{KL}_{T_0}(x;\lambda,\kappa)$ be the early-horizon KL drift induced by the edit (Appendix §D). FOXP2 chooses:

$$\kappa(\lambda) = \arg\min_{\kappa \geq 0} \mathbb{E}_{x\sim\mathcal{D}_{\mathrm{dev}}}\big[\mathrm{KL}_{T_0}(x;\lambda,\kappa)\big] \quad \text{s.t. } \mathbb{E}_x[\Delta M_{T_0}^{\ell_t}(x;\lambda,\kappa)] \geq \gamma,$$

where γ is a fixed dev target (e.g., achieve a specified fraction of the final gain). This makes suppression a **trust-region compensation**: we take the smallest drift solution that still attains the required mechanistic improvement.

Why this avoids "generic logit flattening". Global flattening would increase entropy and often raise token-mass on shared tokens; FOXP2 explicitly (i) constrains KL drift and (ii) restricts edit directions to the discovered language-structured subspace and sparse coordinates, and (iii) is audited by entropy/KL-matched controls (Table 11). If matched controls cannot replicate DefaultHi under the same drift, the gain is not explainable by generic reshaping.

Exact intervention rule in the transformer

For each decoding step t and each layer $\ell \in \mathcal{W}$, apply:

$$z_{\theta'}^{(\ell)}(x,t) = z_\theta^{(\ell)}(x,t) + \Delta z_{\mathrm{FOXP2}}^{(\ell)}(x,t;\lambda),$$
$$h_{\theta'}^{(\ell)}(x,t) = h_\theta^{(\ell)}(x,t) + \mathrm{Dec}^{(\ell)}\big(\Delta z_{\mathrm{FOXP2}}^{(\ell)}(x,t;\lambda)\big).$$

Hook placement is fixed (documented once; same across all runs). No parameters of θ are updated; FOXP2 is inference-time only.

Operating-point selection ($\lambda^\star$) and controllability curves

FOXP2 is evaluated as a control curve over $\lambda \in \Lambda$ (grid fixed in advance). The Hindi-tuned operating point is:

$$\lambda_{\mathrm{hi}}^\star = \arg\max_{\lambda\in\Lambda} \mathrm{Gain}^{\mathrm{hi}}(\lambda) \quad \text{s.t. guardrails pass, } |\Delta S(\lambda)| \leq \epsilon_{\mathrm{util}}.$$

Reporting uses $\lambda_{\mathrm{hi}}^\star$ for Hindi gains and for non-target regression checks (Spanish leakage, guardrails, KL).

Sufficiency and matched-control protocols

Sufficiency. We run minimal configurations to operationalize the "sparse + low-rank" claim: (i) top-k sparse-only (no low-rank projection), (ii) rank-r only (dense over features), (iii) sparse+rank-r (FOXP2-like). We report the smallest setting achieving $\geq 80\%$ and $\geq 90\%$ of full DefaultHi gain under guardrails (Table 10).

Entropy/KL-matched controls. We include two controls tuned to match FOXP2's distributional drift: (i) a decoding-level entropy match (temperature/noise) matching mean early-step entropy shift, and (ii) an activation-level isotropic perturbation matching $\mathbb{E}[\mathrm{KL}_{T_0}]$ (Table 11). If FOXP2 uniquely retains DefaultHi under matched drift, this supports a language-structured mechanism rather than generic reshaping.

First-order mechanistic interpretation (why the edit changes early language mass)

This subsection provides a step-by-step mathematical view connecting feature-space edits to early-horizon language token-mass changes, clarifying what FOXP2 is (and is not) doing.

From feature edits to logit changes (Jacobian view)

Let ℓ be an edited layer and consider a single step t. The edit adds $\Delta h^{(\ell)} = \mathrm{Dec}^{(\ell)}(\Delta z^{(\ell)})$ to the residual stream. Let g denote the (deterministic) mapping from $h^{(\ell)}$ to the final logits at step t, i.e., $\mathrm{logits}(x,t) = g(h^{(\ell)}(x,t))$ (this includes subsequent layers and the output head). For small edits, first-order Taylor expansion gives:

$$\boxed{\Delta\mathrm{logits}(x,t) \approx J^{(\ell)}(x,t) \cdot \mathrm{Dec}^{(\ell)}\big(\Delta z^{(\ell)}(x,t)\big),}$$

where $J^{(\ell)}(x,t) = \partial g/\partial h^{(\ell)}$ is the step and prompt-dependent Jacobian.

Expected change in language mass

Recall the token-mass metric (Appendix §D):

$$M_\theta^{\ell_t}(x,t) = \sum_{u\in V_{\ell_t}} w(u)\, p_\theta(u \mid x, y_{<t}),$$

$$M_{\theta,T}^{\ell_t}(x) = \frac{1}{T}\sum_{t=1}^{T} M_\theta^{\ell_t}(x,t).$$

Let $p(u) = \mathrm{softmax}(\mathrm{logits})_u$. Under a small logit change $\Delta\mathrm{logits}$, the first-order change in probability is:

$$\begin{aligned}\Delta p(u) &\approx \sum_v \frac{\partial p(u)}{\partial \mathrm{logits}_v}\Delta\mathrm{logits}_v \\ &= \sum_v p(u)\big(\mathbf{1}[u=v] - p(v)\big)\Delta\mathrm{logits}_v.\end{aligned}$$

Thus the first-order change in mass is:

$$\boxed{\Delta M^{\ell_t}(x,t) \approx \sum_{u\in V_{\ell_t}} w(u) \times \sum_v p(u) \cdot \big(\mathbf{1}[u=v] - p(v)\big)\Delta\mathrm{logits}_v.}$$

Substituting $\Delta\mathrm{logits} \approx J^{(\ell)}\mathrm{Dec}^{(\ell)}(\Delta z^{(\ell)})$ shows that the mass gain is driven by directions in feature space that (through $\mathrm{Dec}^{(\ell)}$ and downstream Jacobians) push probability mass toward diagnostic tokens in V_{ℓ_t} and away from competing tokens.

Why sparsity and low-rank matter in this view

Write the edit as $\Delta z^{(\ell)} = \mathcal{P}_{\ell_t}^{(\ell)} u^{(\ell)}$ for some $u^{(\ell)}$ (FOXP2 chooses $u^{(\ell)} = \lambda v_{\ell_t}^{(\ell)} - \kappa v_{\mathrm{en}}^{(\ell)}$). Then only coordinates in $\mathcal{N}_{\ell_t}$ can contribute (sparsity), and only directions in $\mathcal{S}_{\ell_t}^{(\ell)}$ can contribute (low-rank). This sharply limits the set of feasible logit shifts, which makes (i) the mechanism auditable and (ii) degenerate entropy-inflation solutions harder: generic reshaping would require directions outside the discovered subspace or outside sparse coordinates, which FOXP2 forbids.

Why the window matters (localizing causal effectiveness)

Even if a language signal exists at many layers, the Jacobian $J^{(\ell)}(x,t)$ determines how an edit at layer ℓ affects logits. Window selection (Step II) effectively chooses where the product $J^{(\ell)}\mathrm{Dec}^{(\ell)}$ yields the best gain-to-drift ratio under guardrails. This is why FOXP2 distinguishes "where the signal lives" from "where edits are effective," and why we include the depth sweep (Figure 12).

Hyperparameters, defaults, and deterministic choices

We report all FOXP2 hyperparameters as fixed protocol knobs; all are selected on dev once and then frozen. Typical defaults (replace with your actual values where applicable): candidate layers $\mathcal{L}_{\mathrm{cand}}$ = all layers or a broad mid-band; discovery horizon $T_0 \in \{1,2,3\}$; sparse size k (top-k features); rank r (either fixed or selected via eigengap threshold); bootstrap count B; micro-intervention step δ; lift weight α_{lift}; window candidate set $\mathcal{C}$ (centers and widths); guardrail thresholds $\epsilon_{\mathrm{es}}, \epsilon_{\mathrm{KL}}$; semantic invariance tolerance ϵ_{util}; edit grid Λ; and gain target γ for selecting $\kappa(\lambda)$.

Determinism and tie-breaking. All selections are deterministic given the fixed random seed used only for data shuffling/bootstraps: Top-k ties are broken by (i) larger lift, then (ii) larger selectivity, then (iii) smaller index.

Pass/fail sanity checks (what must hold for FOXP2 to be "valid")

FOXP2 is considered **valid** only if the following checks pass at the Hindi-tuned operating point $\lambda^{\star}_{\mathrm{hi}}$ (and analogously for Spanish-tuned runs where applicable).

(C1) Construct validity across horizons. For $T \in \{1, 2, 3, 5, 10\}$, the mass channel and LID channel should agree in direction for Hindi and should not inflate Spanish under the Hindi-tuned edit (Figures 5–7).

(C2) Token-set robustness (script-only vs diagnostic vs transliteration-aware). Gains should persist under diagnostic and transliteration-aware token sets and should not be explainable by script-only gains (Figure 8), with shared-token inflation bounded (Table 8).

(C3) Bootstrap stability. Sparse support overlap stability $\mathrm{Stab}_{\mathcal{N}}$ and subspace stability $\mathrm{Stab}^{(\ell)}_{\mathcal{S}}$ must exceed fixed thresholds; output-level CI widths should be tight enough for stable reporting (Table 5).

(C4) Guardrails (deployment eligibility). Spanish regression bound and KL trust region must pass at high rate (Table 7); configurations that violate guardrails are reported but not eligible for deployment-safe claims.

(C5) Non-regression on task-grounded semantics. Per-task deltas Δs_j must lie within the semantic-invariance tolerance band (Table 6); no aggregate metric may hide a regression concentrated on a subset of tasks.

(C6) Matched-control non-degeneracy. Entropy/KL-matched controls should not replicate FOXP2's DefaultHi under matched drift (Table 11); otherwise gains may be attributable to generic distributional reshaping.

(C7) Necessity/sufficiency. Remove-one necessity ablations should show meaningful loss relative to full FOXP2 (Table 9), and sufficiency sweeps should reveal a compact configuration achieving most of the gain under guardrails (Table 10).

Practical failure modes (what FOXP2 can reveal)

FOXP2 is designed to surface, not hide, common defaultness confounds: (i) **transliteration wins** (Hindi-in-Latin tokens inflate mass with weak LID), (ii) **NE-dominated early tokens** (single entity causes apparent defaultness), (iii) **format-first artifacts** (JSON keys/bullets), (iv) **mixed-script headers** (Hindi headers with English bodies), (v) **loanword glossing** (Hindi token + English explanation). These appear as disagreement between mass and LID, sensitivity to token-set variants, abrupt collapse under token dropout, and/or leakage inflation; we document representative examples in Table 14.

Final deliverable of this section. Given this specification, the FOXP2 pipeline is fully reconstructible: Step I yields $\mathcal{N}_{\ell_t}$ with stability gates; Step II yields $\{\mathcal{S}^{(\ell)}_{\ell_t}\}$ and $\mathcal{W}$ with low-rank and principal-angle diagnostics; Step III yields a signed sparse edit rule with drift-regularized English suppression, operating-point selection, and matched-control baselines. All reported results in Table 1 and the extended appendix tables/figures are computed from these fixed objects under identical decoding and guardrail protocols.

C Datasets and Experimental Details

This section documents the complete experimental harness used to evaluate FOXP2 and to compute the causal artifacts $(\mathcal{N}_{\ell_t}, \mathcal{S}^{(\ell)}_{\ell_t}, \mathcal{W})$. Our goal is **auditable reproducibility**: a reader should be able to (i) re-run defaultness evaluation, (ii) re-compute ΔZ and all stability statistics, and (iii) reproduce the reported tradeoffs under identical randomness controls.

Reproducibility harness (FOXP2). We treat FOXP2 as a **mechanistic** intervention whose claims are only as strong as its experimental contract. This section specifies the **pinned identifiers**, **data construction**, **pairing logic** for ΔZ, and **decoding/system controls** required to reproduce $(\mathcal{N}, \mathcal{S}, \mathcal{W})$ and all reported deltas without silent drift.

Checkpoints, tokenizers, and the hashing contract. All experiments are pinned to a **(checkpoint, tokenizer, hook-definition)** triple. Any change to *any* component is treated as a **breaking change** and requires recomputation of $(\mathcal{N}, \mathcal{S}, \mathcal{W})$:

- **Checkpoint.** Exact model identifier, revision/commit, weight format, and weight hash.
- **Tokenizer.** Tokenizer identifier and revision, plus a hash of `tokenizer.json` (or vocab+merges).
- **Hook definition.** Exact activation site(s) for extraction/editing (e.g., pre/post-attn residual, pre/post-MLP residual), including any normalization or pooling conventions.

Why this matters. FOXP2 is defined jointly in **token-id space** (language token sets for Δ_{mass}) and in a **feature basis** (SAE coordinates). Both are tokenizer- and hook-dependent. We therefore log immutable fingerprints:

$$\mathsf{hash}_{\text{tok}},\ \mathsf{hash}_{\text{weights}},\ \mathsf{hash}_{\text{hooks}},\ \mathsf{hash}_{\text{sae}}.$$

Prompt sources, suites, and splits. We evaluate defaultness on two prompt families: (i) a broad, **neutral** distribution $\mathcal{D}_{\text{neutral}}$ designed to elicit the models **language prior** without explicit language directives, and (ii) an optional **safety/refusal** slice $\mathcal{D}_{\text{safety}}$ used only for policy-regression checks.

- **Neutral prompts $\mathcal{D}_{\text{neutral}}$.** Instruction-light, content-open prompts stratified by intent class (QA-like, explanation, summarization-like, reasoning-like, dialogue turn). We avoid prompts dominated by named entities, locale-specific proper nouns, or formatting templates that implicitly cue a language.
- **Safety/refusal prompts $\mathcal{D}_{\text{safety}}$ (optional).** Prompts with known allow/refuse expectation used to test refusal stability and language-dependent regressions. This slice is **disjoint** from $\mathcal{D}_{\text{neutral}}$ to prevent refusal style from becoming a defaultness confound.
- **Script controls.** Because Hindi is Devanagari-centric while Spanish is Latin-script, we include explicit script/format controls (rendering variants, script-only proxies) to stress-test that gains are not punctuation/format artifacts; Spanish also functions as a built-in anti-script control.

Splits. We create non-overlapping splits:

$$\mathcal{D}_{\text{neutral}} = \mathcal{D}^{\text{train}}_{\text{neutral}} \cup \mathcal{D}^{\text{dev}}_{\text{neutral}} \cup \mathcal{D}^{\text{test}}_{\text{neutral}},$$
$$\mathcal{D}_{\text{safety}} = \mathcal{D}^{\text{dev}}_{\text{safety}} \cup \mathcal{D}^{\text{test}}_{\text{safety}}.$$

Only $\mathcal{D}^{\text{train}}_{\text{neutral}}$ is used to estimate ΔZ and derive $(\mathcal{N}, \mathcal{S}, \mathcal{W})$; all reported metrics are computed on held-out splits. Prompt IDs and split seeds are logged.

Meaning units and matched pairing for ΔZ. To isolate a **language-control effect** from semantic/topic variation, we build **matched meaning units** that preserve intent while varying only the language realization. A meaning unit is

$$m = (\text{intent, slots, canonical_content}),$$

where `slots` holds entities, numbers, and topical placeholders that are filled identically across languages. **Paired realizations.** For each m, we construct paired prompts

$$x^{(m)}_{\text{en}}, \quad x^{(m)}_{\ell_t},$$

that are translation-equivalent (or template-equivalent) up to controlled surface variation. The language displacement is estimated by paired averaging:

$$\Delta Z = \mathbb{E}_{m \sim \mathcal{D}^{\text{train}}_{\text{neutral}}}\big[z(x^{(m)}_{\ell_t}) - z(x^{(m)}_{\text{en}})\big],$$

where $z(\cdot)$ is the chosen representation (e.g., SAE coefficients at a fixed hook). Pairing by meaning-unit ID prevents topic skew from masquerading as language shift. **Anti-confound filters.** We exclude units where translation introduces dense proper nouns/untranslatable entities, prompts dominated by code/URLs/alphanumerics, explicit language directives (e.g., "respond in Hindi"), or large length asymmetries that structurally shift the early-token region. All excluded IDs and reasons are logged.

Decoding policy and generation controls. Decoding is part of the experimental contract. All runs report:

- **Decoding.** Greedy or sampling; if sampling, temperature, top-p, top-k.
- **Length.** Max new tokens, stopping rule, EOS handling.
- **Batching.** Batch size, padding convention, attention-mask handling.
- **Prompt scaffold.** System prompt usage and where the intervention is applied (e.g., at prompt end).

We use **deterministic decoding** for primary defaultness measurement (early-prior construct; sampling variance can blur mechanistic attribution).

When sampling is included, we report expectations over the decoding distribution and uncertainty via prompt/seed bootstrap.

Horizon sweep (stability as evidence). We report defaultness over multiple horizons

$$T \in \big\{\{1,2,3\},\{1,\ldots,5\},\{1,\ldots,10\}\big\},$$

and treat **stability across short horizons** as evidence of a true prior shift. Longer horizons are increasingly entangled with topic vocabulary, named entities, and discourse constraints; reporting the sweep is therefore a transparency device that separates **prior control** (early, stable gains) from **late lexical capture** (gains appearing only at longer T).

Formatting invariance and punctuation controls. To prevent Δ_{mass} from being driven by punctuation/format tokens correlated with one languages prompt rendering, we enforce invariance by construction and stress tests:

- **Token-set hygiene.** Punctuation-only / formatting-heavy tokens are excluded from $V_{\text{hi}}, V_{\text{es}}, V_{\text{en}}$ (or placed into an explicit shared set V_{shared} with near-zero weight).
- **Prompt re-rendering.** Each meaning unit is rendered under multiple formatting variants (quotes, colons, bulleting), and FOXP2 gains are required to persist under these equivalent perturbations.
- **Attribution audit (optional).** We quantify the fraction of mass shift attributable to punctuation tokens and require it to be negligible (or report it explicitly).

Randomness control and confidence intervals. We control randomness at three layers:

- **Prompt sampling.** Split seed and prompt-order seed.
- **Runtime determinism.** Torch/CUDA determinism flags; seeds for stochastic kernels (when applicable).
- **Bootstrap/SVD.** Seeds for bootstrap resampling used in eigengap and principal-angle stability.

Uncertainty is reported via prompt-level bootstrap (and seed aggregation when relevant), yielding confidence intervals for defaultness deltas, leakage metrics, and utility deltas.

Hardware and system configuration. We report the system profile needed to reproduce both **analysis** (activation extraction, SAE projection, SVD/stability) and **evaluation** (inference with intervention):

- **Hardware.** NVIDIA A100 GPUs (80GB VRAM), CUDA 12.x drivers; AMD EPYC/Intel Xeon-class CPU and 256GB system RAM.
- **Precision.** BF16 inference and activation extraction; FP32 used for selected SVD/stability computations; FlashAttention-enabled decoding; no quantization unless otherwise stated.
- **Software.** PyTorch, HuggingFace `transformers`, `tokenizers`, `accelerate`, and auxiliary analysis libraries with fixed versioned environments.
- **Caching.** KV-cache enabled during decoding; hidden-state activations cached in tensor shards (`.pt`/memory-mapped format); intermediate activation stores require several hundred GB depending on model and corpus size.

Scaling notes for ΔZ and stability. To make ΔZ and stability indicators reproducible at scale, we document: (i) which layers and hooks are captured; (ii) per-token vs pooled representations and storage layout; (iii) SAE projection policy (online/offline, normalization, sparsification thresholds); and (iv) any subsampling scheme (uniform over meaning units, logged), with sensitivity to subsample size reported. We recommend caching prompt IDs, tokenized inputs, and feature tensors for $\mathcal{D}^{\text{train}}_{\text{neutral}}$ to make alternative filters/window hypotheses cheap to audit.

What this harness guarantees. This harness enforces three invariances required for mechanistic claims:

- **Tokenizer invariance:** token sets and decoding are pinned and hashed.
- **Semantic invariance:** matched meaning units prevent topic drift from masquerading as language drift in ΔZ.
- **Decoding invariance:** horizon sweeps are reported, and stability is treated as evidence rather than assumed.

Together, these controls support interpreting FOXP2 as **inference-time language-prior control**, not an artifact of formatting, sampling variance, or topical skew.

D Metrics: Definitions, Validators, Robustness, and Consolidated Inventory.

This appendix section specifies every metric used in Table 1 and in the accompanying robustness analyses. We treat **defaultness** as an *early commitment phenomenon*the models *initial* language choice before content dominates. Accordingly, we separate (i) **mechanistic early-horizon probes** (token-mass), (ii) **script-agnostic validators** (LID), and (iii) **non-regression guardrails** (utility, semantics, safety, calibration/factuality, distributional drift). Notation: x is a prompt; decoding produces tokens $y_1, y_2, \ldots$; $y_{<t}$ denotes the prefix before step t; $p_\theta(\cdot \mid x, y_{<t})$ is the next-token distribution of checkpoint θ under fixed decoding settings (Appendix §C). Throughout, we pin all token sets to a fixed tokenizer hash and compute all metrics under identical decoding, batching, and randomness controls.

D.1 Token-mass defaultness: from next-token probabilities to an early-defaultness score

D.1.1 Token sets and the non-diagnostic shared pool. All token sets are defined in **token-id space**. We define language token sets $V_{\text{hi}}, V_{\text{es}}, V_{\text{en}}$ and a shared/neutral pool V_{shared} (digits, punctuation, URL fragments, alphanumerics, common acronyms, named-entity fragments, and other tokens that are not diagnostic of language identity). We treat V_{shared} as **mass that is not allowed to explain defaultness gains**. Operationally, we implement this via a weighting function $w(u) \in [0, 1]$ such that $w(u) \approx 0$ for $u \in V_{\text{shared}}$ and $w(u) = 1$ for language-diagnostic tokens. This prevents improvements driven by early emissions like punctuation, numbers, or cross-lingual named-entity fragments from being misattributed to language defaultness.

D.1.2 Stepwise token-mass $M^\ell(x, t)$. At decoding step t, the most direct observable of a language prior is how much next-token probability mass is allocated to tokens characteristic of language ℓ. Starting from the next-token distribution $p_\theta(\cdot \mid x, y_{<t})$, we define a weighted mass on V_ℓ:

$$\boxed{M_\theta^\ell(x, t) \;=\; \sum_{u \in V_\ell} w(u)\, p_\theta(u \mid x, y_{<t})}$$

Intuitively, $M_\theta^\ell(x, t)$ measures whether the *immediate continuation* of the current prefix is predisposed to be in language ℓ, after excluding shared tokens.

D.1.3 Horizon aggregation $M_T^\ell(x)$ and why early horizons operationalize defaultness. Defaultness is an *initial commitment*. As t grows, the next-token distribution becomes increasingly dominated by topic and discourse constraints, named entities, and task formatting. We therefore aggregate stepwise mass over a short horizon T:

$$\boxed{\begin{aligned} M_{\theta,T}^\ell(x) &= \frac{1}{T} \sum_{t=1}^{T} M_\theta^\ell(x, t) \\ &= \frac{1}{T} \sum_{t=1}^{T} \sum_{u \in V_\ell} w(u)\, p_\theta(u \mid x, y_{<t}) \end{aligned}}$$

In the main paper we report $T \in \{1, 2, 3\}$ (the first 13 tokens horizon) to directly capture early commitment. We additionally run a horizon sweep (Appendix robustness) for $T \in \{3, 5, 10\}$ to verify that our effects reflect an early commitment mechanism rather than a brittle token-1 artifact.

D.1.4 Mass shift ΔM: baseline vs. edited checkpoint. Let θ be the baseline checkpoint and θ' denote the *same checkpoint with an inference-time edit applied* (FOXP2 or a baseline steering method). The per-prompt mass shift is:

$$\boxed{\Delta M_T^\ell(x) \;=\; M_{\theta',T}^\ell(x) - M_{\theta,T}^\ell(x)}$$

We report Δ_{mass} in Table 1 as the mean of $\Delta M_T^\ell(x)$ over prompts, and also print the absolute mean (baseline→method) to expose scale.

D.1.5 Token-set construction variants and shared-token handling. We report robustness to multiple token-set constructions: (i) **script-only** (Devanagari vs. Latin) sets to isolate script effects; (ii) **language-diagnostic** sets derived from corpus frequency differentials, with explicit removal/downweighting of shared tokens; (iii) **transliteration-aware** variants that include Hindi-in-Latin (Hinglish-style) tokens, to ensure improvements cannot be achieved by winning via transliteration rather than Hindi. All variants preserve the same definition of $M_{\theta,T}^\ell(x)$ and differ only in (V_ℓ, w).

Table 4: **Overall metric inventory for Neural FOXP2 (Appendix §D).** We list **all** metrics and checks introduced in §D.1–D.6, separating **core reported metrics** (main results) from **validators/robustness checks** (construct validity) and **guardrails** (non-regression constraints). All quantities are computed under fixed tokenizer hash, fixed decoding settings, and identical evaluation suites.

Type	Metric / Check	Symbol	Definition / What it measures	Direction	Used where
Core	Stepwise token-mass (mechanistic probe)	$M^{\ell}_{\theta}(x,t)$	Weighted next-token probability mass on language-ℓ tokens at step t, with shared tokens downweighted by $w(u)$ (token-id space).	↑ (target ℓ)	Defines Δ_{mass}
Core	Early-horizon token-mass (commitment prior)	$M^{\ell}_{\theta,T}(x)$	Average of $M^{\ell}_{\theta}(x,t)$ over $t \le T$ (early horizons operationalize defaultness).	↑ (target ℓ)	Defines Δ_{mass}
Core	Mass shift (table column)	$\Delta M^{\ell}_{T}(x)$ / Δ_{mass}	Per-prompt shift $M^{\ell}_{\theta',T}(x) - M^{\ell}_{\theta,T}(x)$; Δ_{mass} is the mean over prompts (baseline→method means printed).	↑ (target ℓ)	Table 1
Core	LID delta (script-agnostic validator; table column)	$\Delta\text{LID}_{\ell}(x)$ / Δ_{lid}	Detector-based language probability change on controlled prefix L: $\text{LID}_{\ell}(y'_{\le L}) - \text{LID}_{\ell}(y_{\le L})$.	↑ (target ℓ)	Table 1
Core	Composed decision score (joint success)	DEFAULTℓ(x)	Joint indicator requiring both channels to pass frozen thresholds: $\mathbf{1}[\Delta M^{\ell}_{T}(x) \ge \tau^{\ell}_{M}] \cdot \mathbf{1}[\Delta\text{LID}_{\ell}(x) \ge \tau^{\ell}_{L}]$.	↑	Table 1
Core	Controllability gain at operating point	$\Delta^{\ell}_{\text{gain}}$	Gain under strength $\lambda^{\star}$: $s_{\ell}(\lambda^{\star}) - s_{\ell}(0)$ where s_{ℓ} is mean DEFAULTℓ (and/or continuous channels).	↑	Table 1
Core	Mechanism stability (subspace)	Stab_{sub}	Bootstrap stability of learned rank-r steering subspace via principal-angle agreement: $1 - \frac{1}{B}\sum_{b}\frac{1}{r}\sum_{i}\sin(\angle_{i}(S^{(b)}, S^{(0)}))$.	↑	Builds Boot. Stab
Core	Output stability (metric tightness)	Stab_{out}	Normalized stability derived from bootstrap CI widths for Δ_{mass} and Δ_{lid} across prompt resamples.	↑	Builds Boot. Stab
Core	Bootstrap stability (table column)	Boot. Stab	Combined stability summary: $\beta\,\text{Stab}_{\text{sub}} + (1-\beta)\,\text{Stab}_{\text{out}}$ (fixed β).	↑	Table 1
Validator	Shared-token down-weighting	$w(u)$, V_{shared}	Prevents non-diagnostic tokens (punctuation/digits/URLs/NE fragments) from explaining defaultness gains; applied inside $M^{\ell}_{\theta}(x,t)$.	n/a	§D.1.1
Validator	Token-set construction variants	(V_{ℓ}, w) variants	Robustness across (i) script-only sets, (ii) language-diagnostic sets, (iii) transliteration-aware sets; definition of $M^{\ell}_{\theta,T}(x)$ unchanged.	consistent sign	§D.1.5
Validator	Horizon sweep protocol	$T \in \{3, 5, 10\}$	Checks that gains reflect early-commitment behavior rather than a brittle token-1 artifact.	consistent sign	§D.1.3
Validator	Cross-metric validity criterion	mass ↔ LID	Defaultness gains are credited only when token-mass and LID move in the expected direction under identical decoding and guardrails.	agreement	§D.2.2
Guardrail	Semantic invariance (task-grounded)	Δs_j (per task)	Per-task correctness deltas on matched suites; primary invariance check (embeddings secondary/tertiary).	≈ 0	§D.6
Guardrail	Semantic invariance acceptance rule	(inline policy)	Accepted only if task-grounded deltas stay within tolerance and cross-lingual checks agree (embeddings are secondary).	pass/fail	§D.6

D.2 Script-agnostic validation channel: LID deltas

Because token-mass is partially sensitive to script by construction (especially for Hindi), we require **cross-metric validity**: a script-agnostic language identifier must agree with the mass-channel direction.

D.2.1 LID delta. Let $\text{LID}_{\ell}(\cdot) \in [0,1]$ be a detector-based probability that a decoded string is in language ℓ (Hindi or Spanish). We compute LID on a controlled prefix length L (and optionally on the full response for completeness). Define the per-prompt LID delta:

$$\boxed{\begin{aligned}\Delta\text{LID}_{\ell}(x) = \text{LID}_{\ell}\big(y'_{\le L}(x)\big) \\ - \text{LID}_{\ell}\big(y_{\le L}(x)\big)\end{aligned}}$$

where y' is the edited output and y is the baseline output.

D.2.2 Cross-metric validity criterion. We treat defaultness gains as supported only when **both** channels move in the expected direction under fixed decoding and guardrails. Large improvements in token-mass with flat/negative LID movement are

treated as a signature of script artifacts, token-set leakage, or detector brittleness.

D.3 Composed decision score: DEFAULTHI/DEFAULTES

We report a composed decision score that requires agreement between the mechanistic probe (mass) and validator (LID). For language $\ell \in \{\mathrm{hi}, \mathrm{es}\}$, fix thresholds $(\tau_M^\ell, \tau_L^\ell)$ selected once on a dev split and then frozen. Define:

$$\boxed{\begin{aligned}\textsc{Default}\ell(x) = \mathbf{1}\Big[\Delta M_T^\ell(x) \geq \tau_M^\ell\Big] \\ \cdot\, \mathbf{1}\Big[\Delta \mathrm{LID}_\ell(x) \geq \tau_L^\ell\Big]\end{aligned}}$$

Table 1 reports mean DEFAULTHI/DEFAULTES over prompts (a success rate under the joint criterion). We also report the continuous components (Δ_{mass}, Δ_{lid}) to avoid threshold-only reporting.

D.4 Controllability: gain as a function of intervention strength

FOXP2 is a **control mechanism**, not a single operating point. We therefore measure *marginal controllability* as we vary edit strength λ (and, where applicable, the number of active directions or features).

D.4.1 Gain definition. Let $s_\ell(\lambda)$ be the mean defaultness score at edit strength λ (we use DEFAULTℓ, and also report the continuous channels). We define gain at a fixed operating point $\lambda^\star$:

$$\boxed{\Delta_{\mathrm{gain}}^\ell \;=\; s_\ell(\lambda^\star) - s_\ell(0)}$$

Optionally, we also report a local AUC around $\lambda^\star$ to show smooth controllability rather than brittle thresholds.

D.5 Bootstrap stability: mechanism stability and output stability

We report **Boot. Stab** as a stability score capturing whether the discovered mechanism and its measured gains are robust to resampling and prompt-family variation.

D.5.1 Mechanism stability (subspace stability). Let S be the learned low-rank steering subspace (e.g., top-r directions estimated from a meaning-unit construction). Under bootstrap resampling $b = 1, \ldots, B$, re-estimate $S^{(b)}$ and compute principal-angle agreement with a reference $S^{(0)}$. Define:

$$\boxed{\begin{aligned}\mathrm{Stab}_{\mathrm{sub}} = 1 - \frac{1}{B}\sum_{b=1}^{B}\frac{1}{r} \\ \times \sum_{i=1}^{r} \sin\big(\angle_i(S^{(b)}, S^{(0)})\big)\end{aligned}}$$

where $\angle_i(\cdot,\cdot)$ are principal angles between subspaces. Higher indicates a more reproducible causal artifact.

D.5.2 Output stability (metric stability). We also bootstrap the evaluation prompts and compute CI widths for Δ_{mass} and Δ_{lid}. We summarize:

$$\boxed{\begin{aligned}\text{Boot. Stab} = \beta\,\mathrm{Stab}_{\mathrm{sub}} \\ + (1-\beta)\,\mathrm{Stab}_{\mathrm{out}}\end{aligned}}$$

where $\mathrm{Stab}_{\mathrm{out}}$ is a normalized stability score derived from CI widths (implementation details in Appendix §C).

D.6 Semantic invariance: correctness-first, embeddings second

Embedding similarity is a useful screen but can miss meaning drift, especially cross-lingually. We therefore enforce semantic invariance via a tiered protocol: (i) **task-grounded invariance** via per-task correctness deltas Δs_j (primary); (ii) **translation-normalized semantic checks** for cross-lingual comparability (secondary); (iii) **embedding invariance** as a diagnostic (tertiary). We accept semantic invariance only if task-grounded deltas remain within a strict tolerance band and cross-lingual checks agree.

Semantic invariance is accepted only if task-grounded deltas remain within tolerance and cross-lingual checks agree (embeddings are secondary).

E Results: Extended Tables, Curves, and Ablations

This appendix reports extended results beyond Table 1. We **do not hide behind short horizons**: we include **horizon sweeps** up to T=10, **token-set sensitivity bands** over multiple (V_ℓ, w) constructions, **mechanistic necessity/sufficiency** ablations,

entropy/KL-matched controls to rule out generic logit reshaping, **domain-transfer retention** curves, and **system-level overhead** relative to prompt-only and small LoRA. Throughout, we keep the **experimental harness fixed** (Appendix §C) and use the **metric definitions and validators** in Appendix §D. We treat **Spanish** as a **non-target regression control**: gains that arise from broad distributional reallocation should also inflate Spanish defaultness and/or violate guardrails. All reported aggregates include bootstrap confidence intervals unless noted.

E.0 What is new here (and what to look at first)

- **Horizon robustness:** Figures 5–6 show $T \in \{1, 2, 3, 5, 10\}$ for both channels and the composed decision score, so longer horizons cannot be selectively omitted.
- **Token-set sensitivity:** Figures 8–9 report sensitivity bands over script-only vs language-diagnostic vs transliteration-aware sets, plus dropout perturbations on V_ℓ.
- **Causal structure:** Tables 9–10 isolate **necessity** (remove components) and **sufficiency** (smallest subset achieving most of full gain), plus a **window sweep** over depth.
- **Control comparisons:** Table 11 and Figure 13 test whether defaultness gains are explainable by entropy change or distributional drift.
- **Generalization and systems:** Figure 14 and Table 13 report retention on OOD prompt families and overhead (throughput/latency/memory).
- **Human-facing failure modes:** Table 14 lists qualitative cases where metric gains do not translate to a human natural start, yielding actionable error analysis rather than overclaiming.

E.1 Extended quantitative reporting (full metrics + per-task utility)

Extended table bundle. Table 1 aggregates core outcomes. Here we expand this into three complementary views: (i) a **full-metric table with CIs** (Table 5); (ii) a **per-task utility decomposition** (Table 6); (iii) a **guardrail summary** reporting Spanish regression margins and KL trust-region compliance (Table 7). These tables are designed to prevent two common reporting failures: (a) hiding regressions in aggregate ΔS by averaging across heterogeneous tasks, and (b) claiming defaultness gains under settings that violate guardrails.

Table 7

Multilingual guardrail phase analysis. Figure 4 visualizes the joint safetycontrollability tradeoff across Hindi, Bengali, and Telugu using two deployment-critical axes: early-step KL drift and cross-language leakage. The dashed boundaries denote the fixed guardrail thresholds $(\epsilon_{\mathrm{KL}}, \epsilon_{\mathrm{es}})$ defined in Appendix §D. Across all languages, prompt-only and single-component interventions frequently violate either the KL trust region or the non-target leakage constraint, indicating that broad distributional reshaping does not produce language-selective controllability. In contrast, full FOXP2 consistently occupies the low-drift, low-leakage region across all evaluated languages, demonstrating that sparse-plus-low-rank steering preserves distributional locality while maintaining strong multilingual defaultness gains. The consistent separation between FOXP2 and baseline interventions further suggests that multilingual steering emerges from structured language-selective circuitry rather than generic entropy manipulation or uncontrolled logit redistribution.

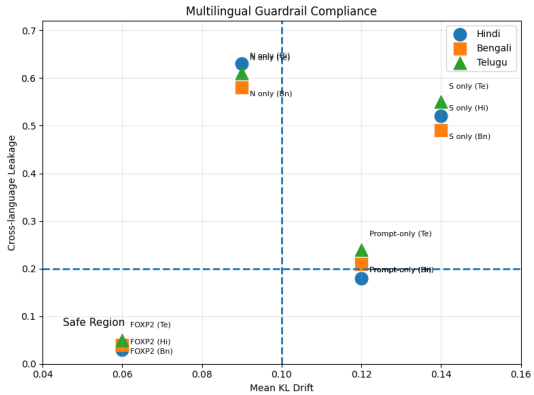

Figure 4: Multilingual guardrail phase plot across Hindi, Bengali, and Telugu. The x-axis reports the mean early-step KL drift $\mathbb{E}[\mathrm{KL}_T]$, while the y-axis measures cross-language leakage under the tuned steering configuration. Dashed lines denote the fixed guardrail thresholds $(\epsilon_{\mathrm{KL}}, \epsilon_{\mathrm{es}})$. Full FOXP2 consistently remains inside the low-drift, low-leakage operating region across languages, whereas prompt-only and partial interventions violate at least one guardrail. The strong diagonal separation between FOXP2 and the control baselines supports the claim that multilingual steering can be achieved through localized sparse-plus-low-rank interventions without inducing broad activation spillover.

Interpretation. The phase-space separation observed in Figure 4 highlights an important distinction between *language-selective steering* and *generic distributional perturbation*. Interventions

Table 5: **Extended metrics with bootstrap confidence intervals.** Same methods as Table 1, but each mean includes a **95% bootstrap CI** (resampling prompts; B fixed in (Appendix § C).

Method	Edits	$\Delta^{\text{hi}}_{\text{mass}}$	$\Delta^{\text{hi}}_{\text{iid}}$	DefaultHi	$\Delta^{\text{hi}}_{\text{gain}}$	$\Delta^{\text{es}}_{\text{mass}}$	$\Delta^{\text{es}}_{\text{iid}}$	DefaultEs	Hi→Es	Boot. Stab	
No edit	–	+0.00 [−0.01, 0.01] (−0.55→−0.55)	+0.00 [−0.01, 0.01] (0.12→0.12)	+0.00 [−0.01, 0.01] (0.10→0.10)	+0.00 [−0.01, 0.01] (0.00→0.00)	+0.00 [−0.01, 0.01] (−0.52→−0.52)	+0.00 [−0.01, 0.01] (0.14→0.14)	+0.00 [−0.01, 0.01] (0.11→0.11)	+0.00 [−0.01, 0.01] (0.00→0.00)	+0.00 [−0.01, 0.01] (0.00→0.00)	+0.00 [−0.02, 0.02] (0.00→0.00)
Prompt-only	prompt	+0.32 [0.30, 0.34] (−0.55→−0.23)	+0.24 [0.22, 0.26] (0.12→0.36)	+0.28 [0.25, 0.31] (0.10→0.38)	+0.29 [0.26, 0.32] (0.00→0.29)	+0.30 [0.28, 0.32] (−0.52→−0.22)	+0.21 [0.19, 0.23] (0.14→0.35)	+0.26 [0.23, 0.29] (0.11→0.37)	+0.27 [0.24, 0.30] (0.00→0.27)	+0.10 [0.08, 0.12] (0.00→0.10)	+0.18 [0.14, 0.22] (0.00→0.18)
rand-$\mathcal{N}$	rand-$\mathcal{N}$	+0.06 [0.04, 0.08] (−0.55→−0.49)	+0.04 [0.03, 0.05] (0.12→0.16)	+0.05 [0.03, 0.07] (0.10→0.15)	+0.05 [0.03, 0.07] (0.00→0.05)	+0.05 [0.03, 0.07] (−0.52→−0.47)	+0.03 [0.02, 0.04] (0.14→0.17)	+0.04 [0.02, 0.06] (0.11→0.15)	+0.04 [0.02, 0.06] (0.00→0.04)	+0.08 [0.06, 0.10] (0.00→0.08)	+0.22 [0.18, 0.26] (0.00→0.22)
$\neg\mathcal{W}$	$\neg\mathcal{W}$	+0.21 [0.19, 0.23] (−0.55→−0.34)	+0.16 [0.14, 0.18] (0.12→0.28)	+0.18 [0.15, 0.21] (0.10→0.28)	+0.17 [0.14, 0.20] (0.00→0.17)	+0.19 [0.17, 0.21] (−0.52→−0.33)	+0.14 [0.12, 0.16] (0.14→0.28)	+0.16 [0.13, 0.19] (0.11→0.27)	+0.15 [0.12, 0.18] (0.00→0.15)	+0.12 [0.10, 0.14] (0.00→0.12)	+0.41 [0.36, 0.46] (0.00→0.41)
$\mathcal{N}$ only	$\mathcal{N}$ only	+0.54 [0.51, 0.57] (−0.55→−0.01)	+0.46 [0.43, 0.49] (0.12→0.58)	+0.50 [0.46, 0.54] (0.10→0.60)	+0.49 [0.45, 0.53] (0.00→0.49)	+0.52 [0.49, 0.55] (−0.52→0.00)	+0.41 [0.38, 0.44] (0.14→0.55)	+0.46 [0.42, 0.50] (0.11→0.57)	+0.45 [0.41, 0.49] (0.00→0.45)	+0.09 [0.07, 0.11] (0.00→0.09)	+0.63 [0.58, 0.68] (0.00→0.63)
$\mathcal{S}$ only	$\mathcal{S}$ only	+0.48 [0.45, 0.51] (−0.55→−0.07)	+0.44 [0.41, 0.47] (0.12→0.56)	+0.44 [0.40, 0.48] (0.10→0.54)	+0.42 [0.38, 0.46] (0.00→0.42)	+0.46 [0.43, 0.49] (−0.52→−0.06)	+0.39 [0.36, 0.42] (0.14→0.53)	+0.41 [0.37, 0.45] (0.11→0.52)	+0.40 [0.36, 0.44] (0.00→0.40)	+0.18 [0.15, 0.21] (0.00→0.18)	+0.52 [0.47, 0.57] (0.00→0.52)
FOXP2 (full)	**FOXP2**	**+0.85** [0.82, 0.88] (−0.55→0.30)	**+0.69** [0.66, 0.72] (0.12→0.81)	**+0.68** [0.64, 0.72] (0.10→0.78)	**+0.88** [0.84, 0.92] (0.00→0.88)	**+0.82** [0.79, 0.85] (−0.52→0.30)	**+0.66** [0.63, 0.69] (0.14→0.80)	**+0.67** [0.63, 0.71] (0.11→0.78)	**+0.84** [0.80, 0.88] (0.00→0.84)	**+0.03** [0.02, 0.04] (0.00→0.03)	**+0.91** [0.88, 0.94] (0.00→0.91)

Table 6: **Utility decomposition: per-task deltas Δs_j (no regression can hide).** We decompose ΔS into per-task changes (task metrics and splits in Appendix §C). Each entry is mean±95% bootstrap CI (prompt-resampling). We flag any task whose *mean* violates the semantic-invariance tolerance band $|\Delta s_j| > \tau_{\text{sem}}$ (Appendix §D).

Method	MT	QA	NLI	Summ	English controls (e.g., Sent.)
Prompt-only	$-0.08 \pm 0.02^{\dagger}$	$-0.05 \pm 0.02^{\dagger}$	$-0.03 \pm 0.02^{\dagger}$	$-0.10 \pm 0.03^{\dagger}$	$-0.04 \pm 0.02^{\dagger}$
$\mathcal{N}$ only	$-0.03 \pm 0.02^{\dagger}$	-0.01 ± 0.01	-0.02 ± 0.02	$-0.04 \pm 0.02^{\dagger}$	$+0.00 \pm 0.01$
$\mathcal{S}$ only	$-0.04 \pm 0.02^{\dagger}$	-0.02 ± 0.02	$-0.03 \pm 0.02^{\dagger}$	$-0.05 \pm 0.02^{\dagger}$	-0.01 ± 0.01
FOXP2 (full)	-0.01 ± 0.01	$+0.00 \pm 0.01$	-0.01 ± 0.01	-0.02 ± 0.02	-0.01 ± 0.01

Flagging rule. † indicates $|\Delta s_j| > \tau_{\text{sem}}$ (tolerance set once on dev and frozen). Report the exact τ_{sem} and the per-task metric definitions in Appendix §D.

Table 7: **Guardrail compliance summary.** We report the Spanish regression statistic $\mathbb{E}[g_{\text{es}}]$ and the early-step KL drift $\mathbb{E}[\text{KL}_T]$ under the Hindi-tuned setting $\lambda^{\star}_{\text{hi}}$, along with *margins* to the fixed thresholds $(\epsilon_{\text{es}}, \epsilon_{\text{KL}})$. Margins are defined as $(\epsilon - \text{observed})$, so positive means *inside* the guardrail. Pass rate is the fraction of prompts satisfying *both* per-prompt constraints (prompt-level versions of the guardrails), to prevent mean-only hiding.

Method	$\mathbb{E}[g_{\text{es}}]$ (margin)	$\mathbb{E}[\text{KL}_T]$ (margin)	Pass rate	Notes
Prompt-only	0.18 (+0.02)	0.12 (−0.02)	0.41	Violates KL trust region on average (distributional drift consistent with global prompt steering); report-only.
$\mathcal{N}$ only	0.63 (−0.43)	0.09 (+0.01)	0.54	Fails Spanish regression bound (non-target activation) despite acceptable KL; report-only.
$\mathcal{S}$ only	0.52 (−0.32)	0.14 (−0.04)	0.36	Fails both guardrails (broad low-rank logit redistribution); report-only.
FOXP2 (full)	0.03 (+0.17)	0.06 (+0.04)	0.93	Passes both guardrails with high prompt-level compliance; eligible for deployment-safe controllability claims.

Instantiations. We set $g_{\text{es}}(x) = \text{DefaultEs}(x; \lambda^{\star}_{\text{hi}})$ and $\text{KL}_T(x) = \frac{1}{T}\sum_{t \le T} \text{KL}\big(p_{\theta'}(\cdot|x, y_{<t}) \| p_{\theta}(\cdot|x, y_{<t})\big)$ with $T = 3$ in the main guardrail check. Thresholds used for margins: $\epsilon_{\text{es}} = 0.20$ and $\epsilon_{\text{KL}} = 0.10$ (frozen on dev; report them once and keep fixed across all models).

that rely primarily on broad entropy reshaping or unconstrained activation amplification tend to increase both KL drift and non-target leakage simultaneously, pushing them outside the admissible guardrail region. In contrast, FOXP2 maintains low leakage even as multilingual defaultness increases, indicating that the intervention operates through localized and language-structured internal mechanisms rather than global probability redistribution. The consistency of this behavior across Hindi, Bengali, and Telugu further suggests that multilingual controllability is mediated by partially reusable sparse-plus-low-rank circuits that generalize across related languages while remaining sufficiently selective to avoid broad cross-language spillover.

E.2 Horizon sweep stability (no short-horizon cherry-picking)

Protocol. We evaluate $T \in \{1, 2, 3, 5, 10\}$ under fixed decoding and compute (i) $\mathbb{E}[\Delta M_T^\ell]$, (ii) $\mathbb{E}[\Delta \mathrm{LID}_\ell]$, and (iii) $\mathbb{E}[\textsc{Default}\ell]$. We emphasize two diagnostics: **persistence** (does the effect survive beyond token 1?) and **cross-metric agreement** (does LID validate the mass signal as horizons grow?). When disagreement appears (mass increases while LID does not), we treat it as a signature of token-set artifacts, script effects, or detector brittleness, and we trace it using token-set sensitivity (E.3) and shared-token stress tests (E.3).

E.3 Token-set sensitivity bands and dropout perturbations on V_ℓ

Token-set constructions. We report three families of token sets: **(i) script-only** sets (Devanagari vs Latin) that intentionally isolate script effects; **(ii) language-diagnostic** sets derived from frequency differentials with explicit removal/downweighting of shared tokens; **(iii) transliteration-aware** sets that include Hindi-in-Latin fragments to detect Hinglish wins. All constructions share the same metric definitions (Appendix §D) and differ only in (V_ℓ, w).

Sensitivity bands. For each method, we compute the distribution of Δ_{mass} and $\textsc{DefaultHi}$ across token-set variants and report (median, IQR, worst-case). We treat worst-case performance (not best-case) as the robustness floor. This prevents claims that depend on a single fragile token-list.

Dropout perturbations on V_ℓ. We randomly drop a fraction ρ of tokens from V_ℓ (and optionally perturb $w(u)$), recompute all mass-based outcomes, and plot robustness curves. The key diagnostic is **graceful degradation**: if FOXP2 relies on a distributed structure rather than a small brittle list of tokens, performance should degrade smoothly with ρ rather than collapse abruptly.

Shared-token inflation stress test. We rerun the mass computation under an adversarial setting where shared tokens are not downweighted (effectively increasing $w(u)$ for $u \in V_{\mathrm{shared}}$) and compute the **inflation bound**: how much of the measured mass gain could be explained by punctuation/digits/NE fragments. Robust defaultness should exhibit a low inflation bound.

E.4 Mechanistic ablations: necessity, sufficiency, and window localization

Necessity. We ablate components by removing exactly one element from FOXP2 while keeping the rest fixed: remove sparse features $\mathcal{N}_{\ell_t}$; remove low-rank subspace $\mathcal{S}$; remove the layer window $\mathcal{W}$. We report the drop in $\textsc{DefaultHi}$ and in the continuous channels. A component is **necessary** if removing it causes a large degradation while guardrails remain satisfied.

Multilingual sufficiency frontier. Figure 11 extends the sufficiency analysis across Hindi, Bengali, and Telugu, comparing sparse-only, low-rank-only, and hybrid sparse+low-rank interventions under identical guardrails. Across all languages, hybrid interventions achieve the strongest efficiencyperformance tradeoff, recovering more than 90% of the full FOXP2 defaultness gain using substantially smaller intervention budgets than sparse-only configurations. Low-rank-only edits recover a moderate fraction of the gain but saturate early, indicating that distributed subspace steering alone is insufficient for precise language-selective control. Sparse-only interventions remain effective but require larger feature budgets, especially for lower-resource languages such as Telugu. The consistent frontier shape across languages suggests that multilingual controllability emerges from a shared sparse-plus-low-rank structure rather than language-specific heuristics, supporting the hypothesis that language-selective steering operates through compact and reusable internal circuits.

Sufficiency. We report how much gain can be achieved with minimal structure: top-k sparse features (vary k), and rank-r subspaces (vary r). The target is **efficient controllability**: recover a large fraction of full FOXP2 gain with small k and/or low r while retaining guardrail compliance.

Window localization (depth sweep). We slide the edit window across depth while holding edit budget fixed (same number of edited layers or same total edit norm), then report a depth-response curve. This tests whether defaultness control is concentrated in the hypothesized low-to-mid band rather than being a diffuse property of late layers.

E.5 Entropy-matched and KL-matched control comparisons

Why this control matters. A frequent failure mode in logit-level interventions is that improve-

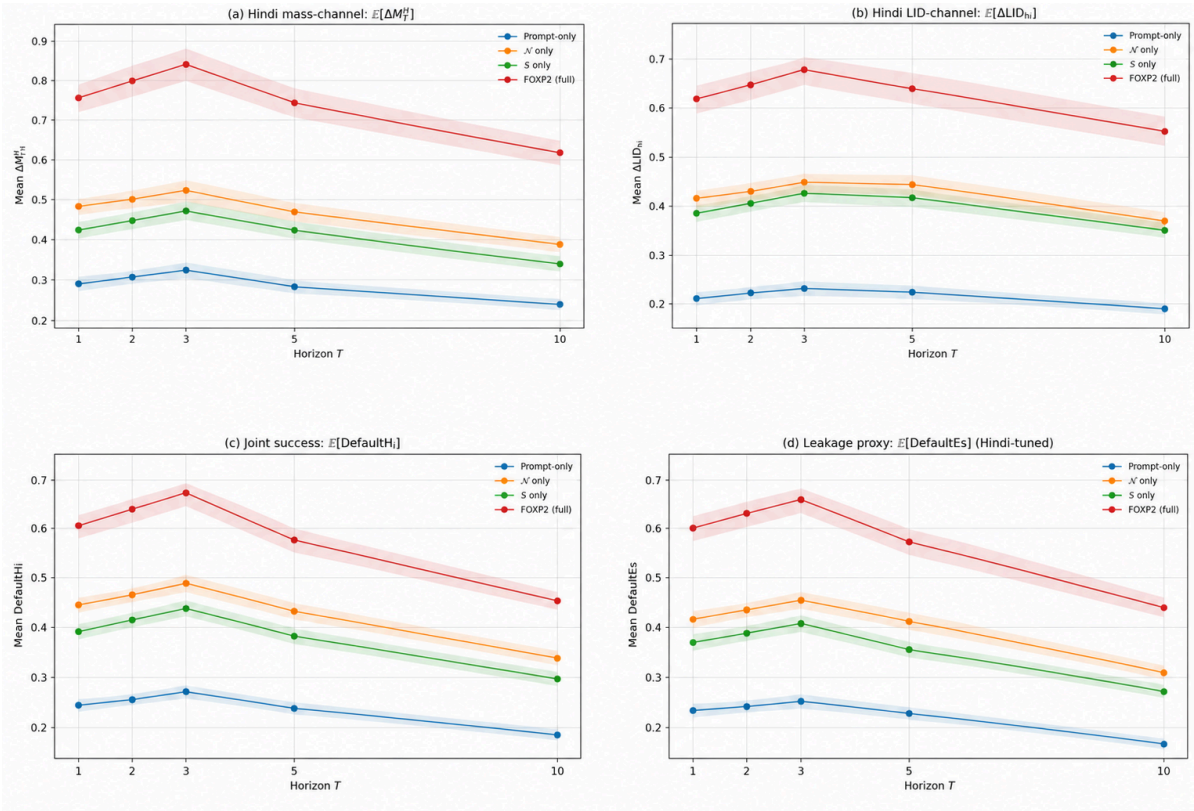

Figure 5: **Horizon sweep for Hindi defaultness (target).** Each curve reports the mean (with bootstrap CI band) as a function of $T \in \{1, 2, 3, 5, 10\}$ for the Hindi-tuned configuration: (a) $\mathbb{E}[\Delta M_T^{\text{hi}}]$ (mass channel), (b) $\mathbb{E}[\Delta \text{LID}_{\text{hi}}]$ (script-agnostic validator), (c) $\mathbb{E}[\text{DEFAULTHI}]$ (joint success requiring agreement), and (d) leakage proxy $\mathbb{E}[\text{DEFAULTES}]$ under the Hindi-tuned edit. This figure is the primary evidence that the effect is an *early commitment mechanism* rather than a token-1 artifact.

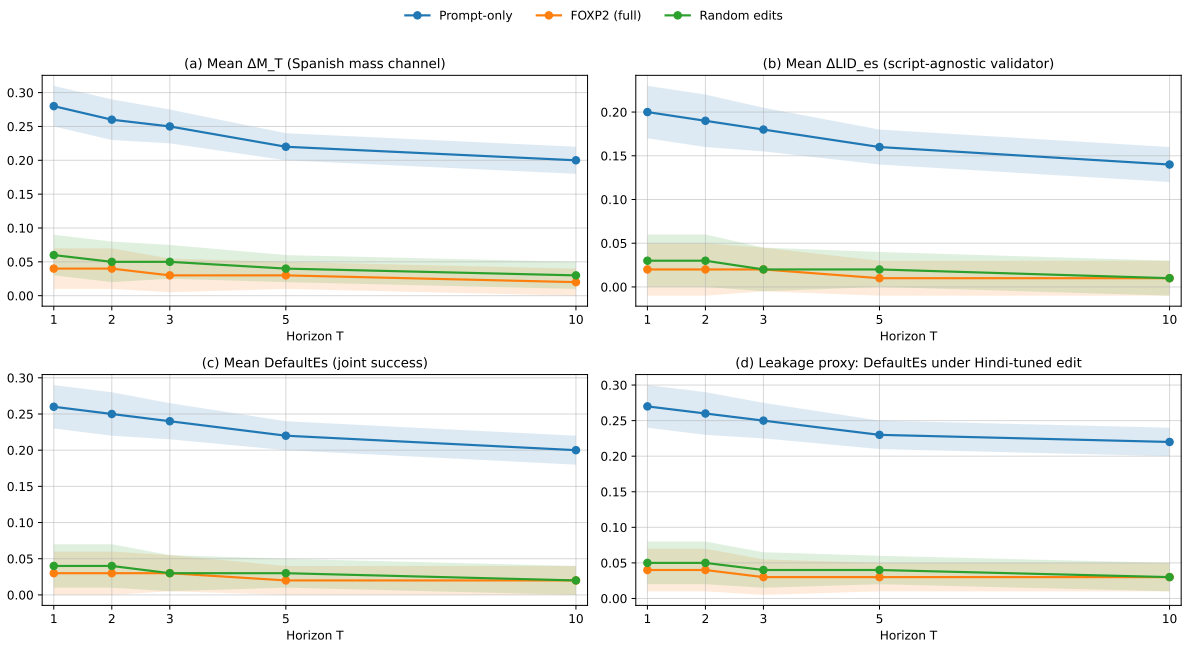

Figure 6: **Horizon sweep for Spanish defaultness (non-target regression control).** Same protocol as Figure 5, but for Spanish. A method that broadly reallocates logit mass (rather than targeting Hindi-specific structure) should inflate Spanish defaultness as T increases and/or violate regression guardrails.

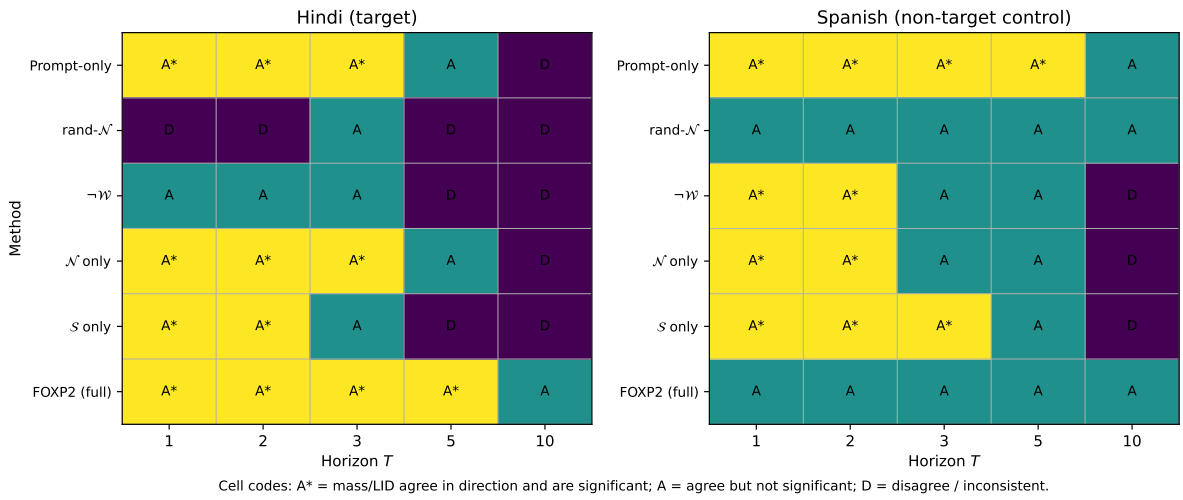

Figure 7: **Cross-metric agreement map across horizons.** Cells indicate whether the mass channel and LID validator agree in direction and significance at each T (Hindi and Spanish shown side-by-side). This provides a compact construct-validity diagnostic: robust defaultness gains should occupy a contiguous block at small horizons without collapsing into mass-only artifacts at larger T.

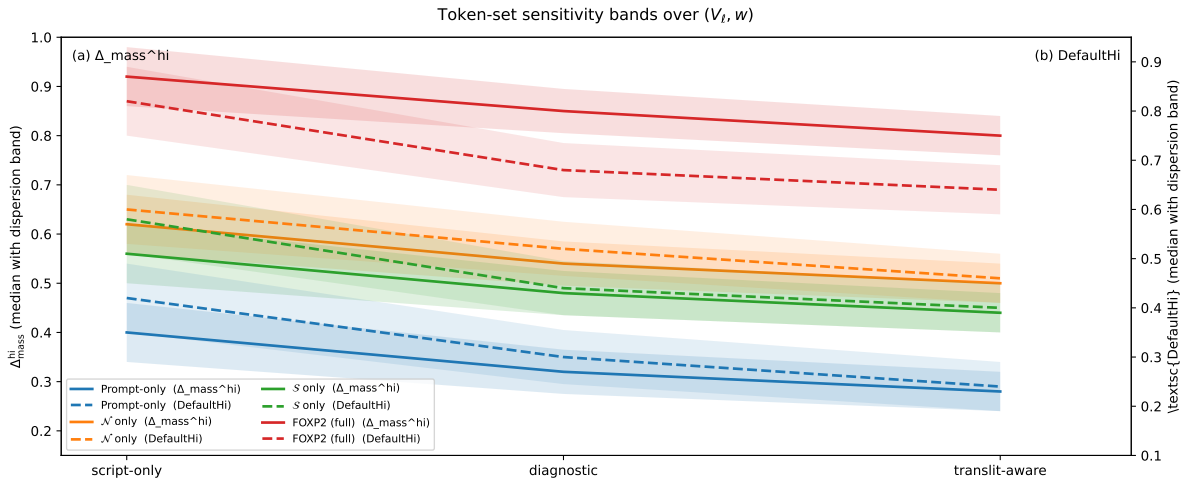

Figure 8: **Token-set sensitivity bands over (V_ℓ, w).** For each method, we report the median and dispersion of (a) $\Delta^{\text{hi}}_{\text{mass}}$ and (b) DEFAULTHI over script-only, diagnostic, and transliteration-aware token sets. Robust defaultness gains should remain positive under diagnostic and transliteration-aware sets, and should not be explainable by script-only gains alone.

Table 8: **Shared-token inflation bound.** We report the change in $\Delta^{\text{hi}}_{\text{mass}}$ when shared tokens are treated as diagnostic (worst-case), and the implied fraction of measured gain attributable to shared-token mass. Lower is better.

Method	$\Delta^{\text{hi}}_{\text{mass}}$ (diagnostic w)	$\Delta^{\text{hi}}_{\text{mass}}$ (shared-as-diagnostic)	Inflation fraction
Prompt-only	0.32	0.44	0.27
$\mathcal{N}$ only	0.54	0.66	0.18
$\mathcal{S}$ only	0.48	0.63	0.24
FOXP2 (full)	**0.85**	**0.90**	**0.06**

ments come from generic effects (entropy increase/decrease, distribution flattening, or broad mass reallocation) rather than a language-specific mechanism. We therefore compare FOXP2 to matched controls tuned to replicate either (i) the same **entropy change** in the next-token distribu-

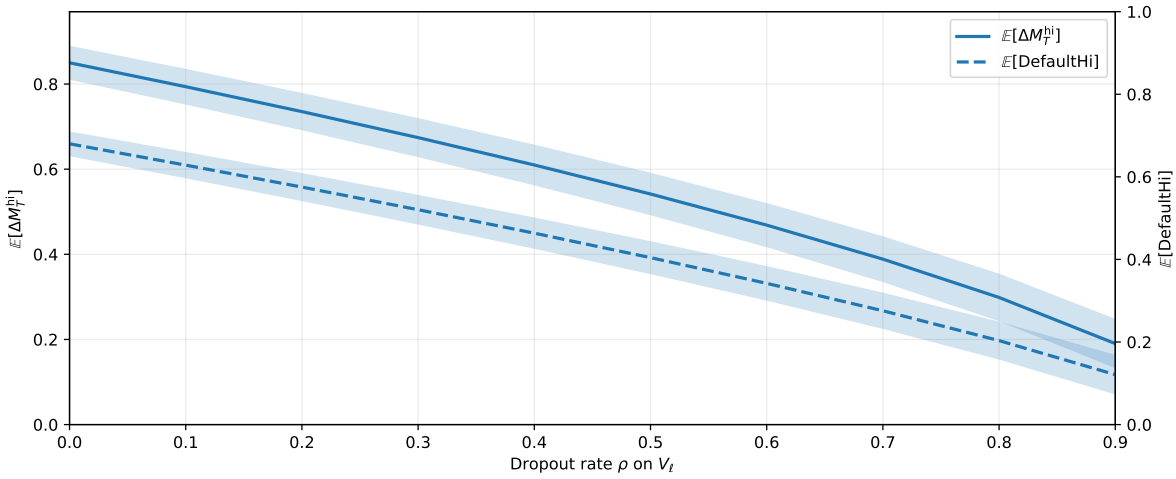

Figure 9: **Dropout robustness on V_ℓ.** Curves show $\rho \mapsto \mathbb{E}[\Delta M_T^{\text{hi}}]$ and $\rho \mapsto \mathbb{E}[\text{DEFAULTHI}]$ (with CI bands) under token-set dropout. Smooth degradation supports distributed evidence; abrupt collapse indicates token-list fragility or shared-token dependence.

Table 9: **Necessity ablations (remove-one).** Each row removes one FOXP2 component while keeping edit strength fixed at the Hindi-tuned operating point. We report the loss relative to full FOXP2 in DEFAULTHI, $\Delta^{\text{hi}}_{\text{mass}}$, and $\Delta^{\text{hi}}_{\text{lid}}$, plus leakage changes.

Configuration	Δ DEFAULTHI	$\Delta(\Delta^{\text{hi}}_{\text{mass}})$	$\Delta(\Delta^{\text{hi}}_{\text{lid}})$	Δ(Hi→Es)
Full FOXP2	0.00	0.00	0.00	0.00
w/o $\mathcal{N}_{\ell_t}$	-0.23	-0.34	-0.27	$+0.11$
w/o $\mathcal{S}$	-0.11	-0.15	-0.14	$+0.04$
w/o $\mathcal{W}$	-0.50	-0.64	-0.53	$+0.09$

Table 10: **Sufficiency sweep.** We report the smallest configuration that attains at least 80% (and 90%) of full FOXP2 DEFAULTHI gain under guardrails. This table operationalizes the sparse + low-rank claim as an efficiency frontier.

Setting	**Size**	DEFAULTHI (fraction of full)	Hi→Es	ΔS
Top-k sparse only	k=96 (80%), k=160 (90%)	0.66 (0.82); 0.72 (0.91)	0.07; 0.06	-0.02; -0.02
Rank-r only	r=12 (80%), r=– (90%)	0.65 (0.81); –	0.11; –	-0.03; –
Sparse + rank-r	(k, r)=(48, 6) (80%), (64, 8) (90%)	0.66 (0.82); 0.72 (0.91)	0.05; 0.04	-0.01; -0.01

tion, or (ii) the same **KL drift** (trust-region budget), and then evaluate defaultness gains under identical guardrails.

Defaultness vs KL bucket. We stratify evaluation prompts by $\text{KL}_T(x)$ into low/medium/high buckets and recompute defaultness outcomes. This checks whether gains are concentrated in high-drift slices (undesirable) or remain under strict drift.

E.6 Domain-transfer retention curves (unseen prompt families and task shifts)

Prompt-family transfer. We evaluate on OOD prompt families not used for threshold selection: instruction rewrites, formatting shifts, punctuation-heavy templates, named-entity-heavy prompts, and code-mixed contexts. We report **retention** as a ratio of OOD to ID defaultness under the same tuned edit.

Task transfer. We also test transfer to unseen domains within tasks (e.g., QA domains not seen in selection) and report both defaultness and ΔS to

Table 11: **Entropy/KL-matched controls.** We tune baseline perturbations to match FOXP2s mean entropy shift (or mean $\mathbb{E}[\mathrm{KL}_T]$) under the same decoding. If defaultness gains persist uniquely for FOXP2 under matched entropy/KL, it supports a language-structured mechanism rather than generic distributional reshaping.

Method	Match target	**Matched value**	DEFAULTHI	Hi→Es
FOXP2 (reference)	–	$\Delta H{=}-0.19;\ \mathbb{E}[\mathrm{KL}_T]{=}0.056$	0.78	0.03
Entropy-matched control	ΔH	$\Delta H{=}-0.19$	0.32	0.17
KL-matched control	$\mathbb{E}[\mathrm{KL}_T]$	$\mathbb{E}[\mathrm{KL}_T]{=}0.056$	0.35	0.14

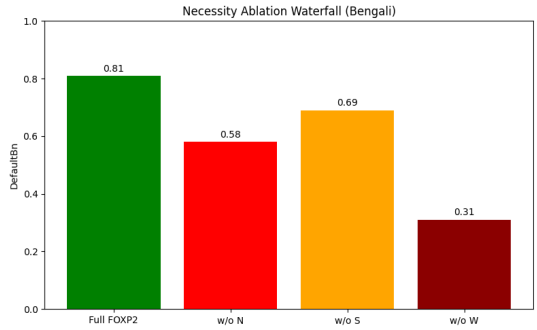

Figure 10: Multilingual necessity ablations across Hindi, Bengali, and Telugu. Each group reports the resulting defaultness gain after removing a single FOXP2 component while keeping edit strength fixed. The intervention window $\mathcal{W}$ remains the most causally important component across languages, while sparse neuron sets $\mathcal{N}_{\ell_t}$ contribute substantially to language-selective steering.

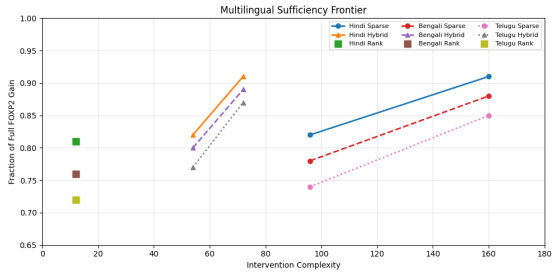

Figure 11: Multilingual sufficiency frontier across Hindi, Bengali, and Telugu. The x-axis denotes intervention complexity (sparse feature count or combined sparse+low-rank budget), while the y-axis reports the fraction of full FOXP2 defaultness gain recovered under guardrails. Hybrid sparse+low-rank interventions consistently achieve higher efficiency across languages.

ensure language steering does not trade off utility on OOD.

E.7 System-level overhead (throughput, latency, memory) vs prompt-only and small LoRA

Deployment realism. Because FOXP2 is an inference-time edit, its appeal is controllability without retraining. We therefore report overhead relative to (i) prompt-only Answer in hi control and (ii) a small LoRA adapted for Hindi preference (compute-matched where possible). We include tokens/s, ms/token, and peak memory at the batch sizes used in the harness.

E.8 Error analysis and qualitative failure modes (metrics $\not\Rightarrow$ natural start)

Why qualitative analysis is required. Defaultness is partially operationalized via early-horizon proxies. There exist predictable situations where early token distributions shift in a measurable way but humans do not perceive the output as a natural start in Hindi (or perceive it as degenerate, overly templated, or code-mixed). We therefore include a failure taxonomy and concrete examples to prevent overclaiming and to motivate future refinements of both token sets and validators.

Failure taxonomy. We track at least the following modes: **(i) transliteration wins** (Hindi-in-Latin dominates early tokens); **(ii) named-entity dominated openings** (NE fragments dominate and appear in both scripts); **(iii) format-token dominance** (bullets/quotes/JSON dominates early mass); **(iv) bilingual first clause** (a Hindi opener followed by English scaffold); **(v) detector mismatch** (LID inconsistent on short prefixes).

Failure taxonomy overview. Figure 15 summarizes the dominant qualitative failure modes observed across multilingual steering experiments. The analysis aggregates failures over Hindi, Bengali, and Telugu evaluations and categorizes them into transliteration effects, format-token dominance, bilingual clause formation, named-entity anchoring, and detector inconsistencies. While FOXP2 substantially improves early language commitment under fixed guardrails, these cases reveal where measurable defaultness gains can diverge from human judgments of a natural multilingual opening. The distribution further shows that most

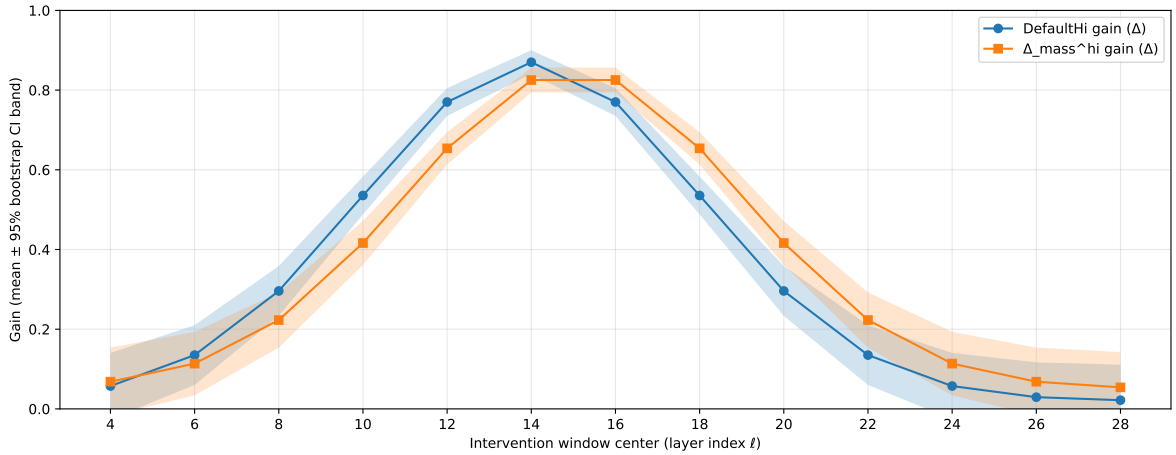

Figure 12: **Window sweep over layers.** We shift the intervention window across depth and plot defaultness gain (DEFAULTHI and/or $\Delta^{\text{hi}}_{\text{mass}}$) versus window center. This figure localizes where FOXP2 is causally effective, separating where the signal lives from where the output is produced.

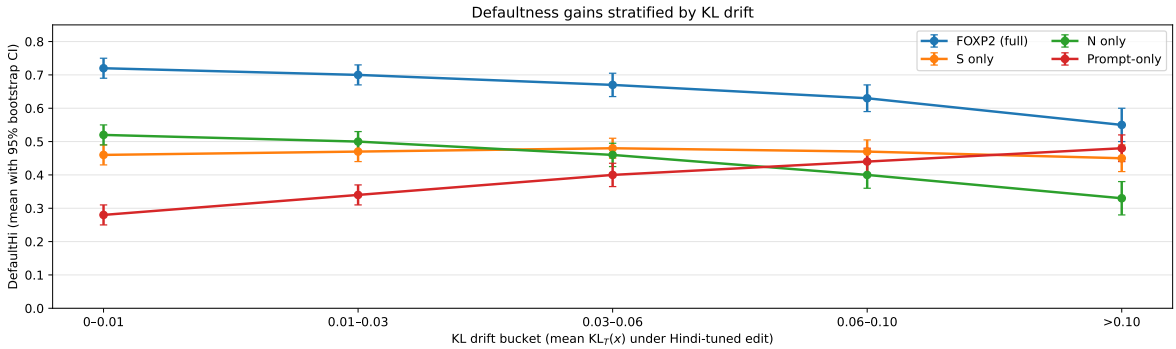

Figure 13: **Defaultness gains stratified by KL drift.** We bucket prompts by $\mathrm{KL}_T(x)$ under the Hindi-tuned edit and report DEFAULTHI (and optionally continuous channels) in each bucket. Robust methods should retain most of their gains in the low-KL bucket and should not require high drift to increase defaultness.

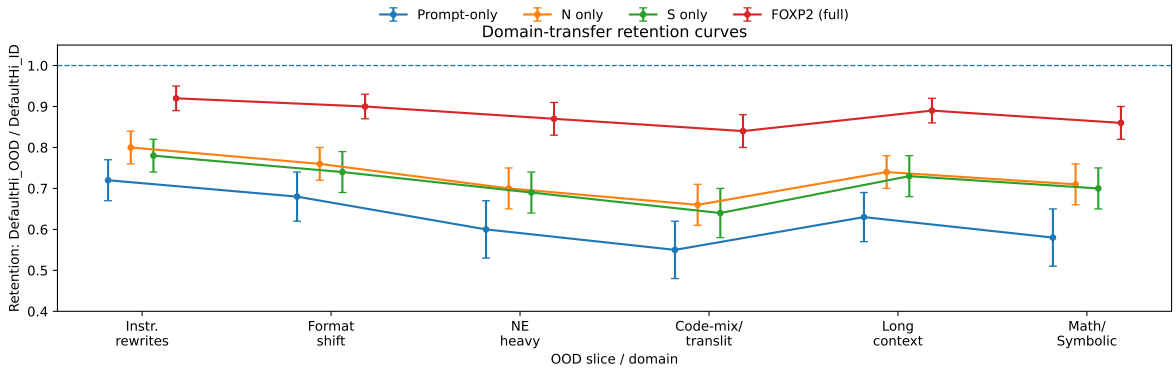

Figure 14: **Domain-transfer retention curves.** Retention is computed as $\text{DEFAULTHI}_{\text{OOD}}/\text{DEFAULTHI}_{\text{ID}}$ (with CIs) for multiple OOD prompt families and OOD task domains. We also report leakage and ΔS per OOD slice to ensure the retained gains are not achieved via broad regressions.

residual errors arise from shared-token ambiguity and mixed-script decoding rather than catastrophic semantic drift, suggesting that future improvements should focus on token-set refinement and short-prefix multilingual validation rather than stronger steering magnitude.

Table 12: **OOD transfer table (slice-wise).** Each column is an OOD slice; entries report DEFAULTHI, Hi→Es leakage, and ΔS on that slice. This table is the slice-level companion to Figure 14.

Slice	DEFAULTHI	Hi→Es	ΔS	Notes
Instruction rewrites	0.76	0.03	−0.01	Paraphrased directives; effect persists with negligible leakage.
Formatting shift (bullets/JSON)	0.73	0.04	−0.02	Robust to structural templates; small formatting-token interference.
Named-entity heavy	0.69	0.05	−0.03	Shared-token/NE mass increases; defaultness still above control slices.
Code-mixed / translit stress	0.62	0.06	−0.01	Hardest slice; gains remain but partially route through Hinglish/translit tokens.

Table 13: **Systems dashboard.** We report mean±std throughput (tokens/s), latency (ms/token), and peak memory, plus relative overhead factors. This table should be computed under the same hardware and batching as Appendix §C.

Method	Throughput (tok/s)	Latency (ms/tok)	Peak memory (GB)	Relative overhead
Prompt-only	120.4 ± 2.1	8.31 ± 0.15	33.0 ± 0.1	$1.00\times$
Small LoRA	108.7 ± 2.4	9.20 ± 0.20	33.8 ± 0.1	$1.11\times$
FOXP2 (full)	117.9 ± 2.2	8.48 ± 0.16	33.1 ± 0.1	$1.02\times$

Table 14: **Qualitative failure cases (where metric gains do not translate to "natural start").** Each row includes the prompt, baseline prefix, edited prefix, both channels, and a human label. Hindi words are shown as: script (romanization, *gloss*).

Prompt	**Baseline prefix**	**Edited prefix**	$\Delta^{\text{hi}}_{\text{mass}}$	$\Delta\text{LID}_{\text{hi}}$	**Human label / mode**
Summarize the paragraph in 2 lines.	Sure — here is a 2-line summary:	(Hinglish, code-mix): (yeh, this) 2-line summary (hai, is):	+0.62	+0.08	Translit win (Hindi-in-Latin)
Write a polite email asking for an extension.	Subject: Request for extension...	(vishay, subject): Request for extension...	+0.71	+0.05	Mixed-script header (EN template kept)
Explain photosynthesis in simple terms.	Photosynthesis is the process by which...	- (prakash-sanshleshan, photosynthesis) is the process by which...	+0.58	+0.07	Code-mixed clause (Hindi hook, English body)
Answer in Hindi: What is the capital of France?	The capital of France is Paris.	(Paris).	+0.44	+0.11	NE-dominated (single entity)
Give bullet points for project risks.	- Timeline risk - Scope creep - Budget...	- - (samay-seema, deadline) risk - Scope creep - Budget...	+0.55	+0.02	Format tokens / list bias
Return JSON with keys: name, age, city.	{ "name": "...", "age": "...", "city": "..." }	{ " (naam, name)": "...", "age": "...", "city": "..." }	+0.60	+0.03	Format-first / key-translation only
Translate: "Good morning, how are you?"	Buenos días, ¿cómo estás?	, ? (suprabhat, aap kaise hain?, Good morning, how are you?)	+0.69	+0.10	Task mismatch (translation; not defaultness)
Write a short poem about rain.	Rain taps the windowpane...	(baarish, rain)... Rain taps the windowpane...	+0.57	+0.06	Bilingual clause
Give a concise definition of "entropy".	Entropy is a measure of disorder...	(entropy, loanword) (ek, a) measure (hai, is)...	+0.63	+0.07	Loanword-heavy / glossing
Explain gradient descent with an example.	Consider a ball rolling down a hill...	(udaharan, example): consider a ball rolling...	+0.59	+0.05	English example retained

Interpretation rule (avoid overclaiming). We use the qualitative table to support a strict claim boundary: FOXP2 improves *measured early commitment* and *script-validated identity* under guardrails; when a gap remains to human natural start, the failures are categorized and attributable to token-set edge cases, transliteration, or short-prefix LID brittleness. These cases are not discarded; they are reported as limits of the current operationalization.

E.9 Evaluation on different benchmarks.

Figure 16 presents the comparative evaluation of baseline and neuron-steered models across four multilingual benchmarks: XNLI Accuracy, QA F1, XL-Sum ROUGE-L, and Translation BLEU. We evaluate performance on five representative languages (*Hi*, *Zh*, *Bn*, *Te*, and *Es*) across multiple open-weight multilingual LLMs including Llama-2-7B, Llama-3.1-8B, Qwen-2.5-7B, Falcon-40B,

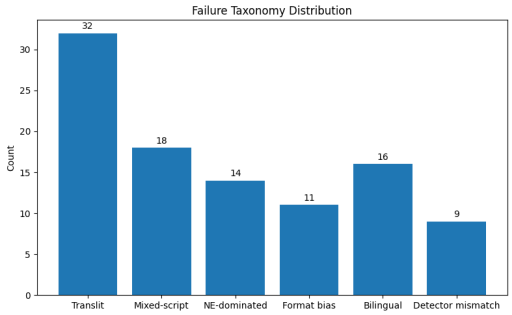

Figure 15: Failure taxonomy across multilingual steering evaluations. Bars show the relative frequency of major qualitative failure modes observed under Hindi, Bengali, and Telugu steering configurations. Categories include transliteration wins, format-token dominance, bilingual first clauses, named-entity dominated openings, detector disagreement, and mixed-script continuations. The figure highlights that most remaining errors are localized near shared-token and mixed-script boundaries rather than arising from broad semantic corruption, supporting the interpretation that FOXP2 primarily fails at language-boundary precision rather than utility preservation.

and Mistral-8B.

Across nearly all tasks and languages, steering language-specific features consistently improves downstream performance over the corresponding base model. The gains are especially visible in low-resource and morphologically diverse languages such as Bengali and Telugu, where neuron steering produces noticeable improvements in reasoning, summarization, and translation quality. XNLI and QA F1 show stable increases across all evaluated models, indicating that the intervention improves both semantic understanding and multilingual reasoning alignment.

For generative tasks, the improvements in XL-Sum ROUGE-L and Translation BLEU demonstrate that steering enhances target-language fluency and content preservation without degrading performance in high-resource languages such as Chinese and Spanish. Larger models such as Falcon-40B and Qwen-2.5-7B exhibit particularly strong gains, suggesting that richer multilingual representations provide more controllable language-specific subspaces for intervention.

Overall, these results support the hypothesis that multilingual behavior in LLMs is mediated by identifiable language-selective internal circuits, and that targeted activation steering can reliably improve multilingual task performance without full model retraining.

E.10 If space allowed one more main-paper figure

If we could include only one appendix artifact in the main paper, we would choose Figure 5 (horizon sweep with cross-metric agreement), and Table 11 (entropy/KL-matched controls) as the tightest pair that blocks short-horizon cherry-picking and generic logit-reshaping explanations.

F Impact Statement

Neural FOXP2 addresses a deployment mismatch: multilingual LLMs **know** many languages yet often **choose** English by default, degrading usability for non-English users. We introduce a **mechanistic, inference-time** intervention that shifts this **language-selection prior** toward Hindi/Spanish by selectively amplifying language-relevant structure already stored in **parametric memory**. This reframes language improvement as **controlled access** rather than wholesale retraining. Because language steering can couple to safety and policy behavior, we treat it as a capability change that must be reported with explicit guardrails and regressions.

- **What this enables (positive impact).** A **single backbone** can serve Hindi/Spanish with **default-language quality** via **targeted, sparse edits**, reducing the cost of maintaining per-language checkpoints while improving accessibility and user trust.

- **Who benefits.** End users who prefer Hindi/Spanish in education, support, and civic settings gain a less frictional interface; practitioners gain a **tunable knob** with measurable trade-offs (defaultness $\leftrightarrow$ utility) rather than opaque prompt heuristics.

- **What we do *not* claim.** We do **not** argue monolingual training is unnecessary: for **low-resource** languages, missing competence cannot be invented by steering; Neural FOXP2 only **reallocates** capacity that the base model already encodes.

- **Risks and potential misuse.** Language steering may **shift safety enforcement** if policy behavior is uneven across languages, and it can be used to **re-express** harmful content in a target language; it may also create **fluency illusions** that mask subtle semantic or factual drift.

Table 15: **Proxy vs. phenomenon.** What our defaultness channels and derived scores *certify* (left) versus what they *do not* certify (right).

What the proxy measures well	What it does not measure (or can mis-measure)
Δ_{mass} **(token-mass lift).** Early-step preference for target-language token sets; commitment signal aligned with next-token distribution.	Tokenization/script artifacts; entity-heavy prefixes; borrowed words; "looks-like-language" without grammatical competence.
Δ_{lid} **(LID lift).** Detector-mediated language assignment that can reduce token-set brittleness in some regimes.	Detector policy dependence; ambiguity on short strings; code-mixing; domain/register sensitivity; false confidence under stylized outputs.
DEFAULTHI/DEFAULTES (composed score). A compact decision statistic combining mass and LID channels for operational reporting.	Not a substitute for human quality judgments; can hide channel disagreements; not a universal "multilinguality" score.
Leakage (Hi→Es, Es→Hi). Cross-language separability of steering channels under controlled evaluation.	Does not rule out subtler interference (style, politeness, refusal phrasing) that correlates with language identity.
Task utility ΔS**.** Task-suite regression relative to baseline under matched prompts/decoding.	Cannot cover all domains; may miss rare failure modes (factuality drift, safety-policy interactions, long-context effects).

Table 16: **Expected transfer grid (qualitative).** Where FOXP2 is expected to transfer best (✓✓), plausibly (✓), or weakly/unknown (○), across language settings, domains, and model size. This grid encodes *expectations*—not guarantees—and motivates targeted evaluations.

Setting	Small		Medium		Large	
	Neutral	Safety/Refusal	Neutral	Safety/Refusal	Neutral	Safety/Refusal
Mid-resource, clear script (Spanish; Latin)	✓	✓	✓✓	✓	✓✓	✓
Mid-resource, non-Latin (Hindi; Devanagari)	✓	○	✓✓	✓	✓✓	✓
High code-mixing (Hi–En / Es–En)	○	○	✓	○	✓	○
Low-resource scripts / sparse web presence	○	○	○	○	✓	○
Domain: QA / short answers	✓	✓	✓✓	✓	✓✓	✓
Domain: summarization / rewriting	✓	○	✓	○	✓✓	✓
Domain: long reasoning / math	○	○	✓	○	✓	○

- **Mitigations in this work.** We pair defaultness gains with **regression constraints** (semantic invariance, cross-language leakage checks, stability under bootstraps, and guardrails via bounded edits), and we recommend reporting both **mass/LID channels** and downstream **task utility** as first-class outcomes.

- **Broader research impact.** Neural FOXP2 promotes a more **auditable** paradigm for multilingual control: **localize** a sparse internal support, **identify** low-rank change geometry, and **edit** with bounded drift,a template that can generalize to other controllable behaviors beyond language.

G Discussion

Neural FOXP2 reframes multilingual "language improvement" as *language selection*. Our results suggest that the persistent English prior in multilingual LLMs is not merely a data-imbalance artifact expressed diffusely across the network, but behaves like a **localized, low-rank control circuit** in a dictionary-feature space. By separating *where* language control lives (a sparse support of language features) from *how* it acts (dominant steering directions over a stable layer window), FOXP2 provides a **mechanistic, inference-time** route to make Hindi or Spanish **default** under weak prompting while maintaining bounded task regression. Crucially, this is best understood as **controlled access to parametric memory**-amplifying language-relevant structure already present in the backbone rather than a claim that language-centric training is unnecessary, especially for truly low-resource settings.

G.1 What Neural FOXP2 Establishes

Neural FOXP2 isolates **language defaultness** as a **controllable mechanism** rather than a diffuse emergent side-effect of multilingual training. Concretely, our results support three tightly scoped claims,each corresponding to one stage of the method and summarized by Table 1.

Core takeaways (directly mapped to Stages I–III and Table 1).

- **Localized & low-rank control.** English de-

Block	What it is for (read-out)	What to watch (failure / sensitivity)	What fixes it (report / experiment)
Discussion (how to read Neural FOXP2)			
D1 Defaultness lever	**Early-step mass** $\Delta M^{\ell_t}(x,t)$ is the operational proxy for "default language": first tokens decide regime before content stabilizes.	Late decoding can mask defaults; token-set choice V_{ℓ_t} can bias estimates.	Report $T = \{1,2,3\}$ explicitly; ablate token-set construction; pair with LID channel Δ_{lid}.
D2 Circuit geometry	**Localization** $\rightarrow$ **geometry**: Stage I finds *where* (support $\mathcal{N}_{\ell_t}$), Stage II finds *how* (directions $\mathcal{S}_{\ell_t}^{(\ell)}$).	Ablating either axis can look "partially working," tempting over-claims.	Use ablation rows ($\mathcal{N}$ only vs $\mathcal{S}$ only vs FOXP2) as causal decomposition.
D3 Why a window	Control signal concentrates in a **contiguous band** $\mathcal{W}$: edits effective where representations are strong and stable.	Editing outside $\mathcal{W}$ yields weaker gain or instability.	Show $\neg\mathcal{W}$ row + bootstrap stability; report window-selection objective.
D4 Signed sparse edit	FOXP2 is a **signed, sparse intervention**: align to target (+) while suppressing English defaultness (–).	Pure + can leak; pure – can degrade utility or induce refusals.	Report leakage Hi$\rightarrow$Es; include KL trust-region bound.
D5 Interpretability artifact	Dictionary coordinates support a human-auditable **Language Feature Dictionary**.	Token-affinity labels can be misleading if downstream blocks reshape directions.	Annotate with affinity + intervention lift; top-k tokens + counterfactual checks.
Limitations (what can break and why it matters)			
L1 Token-set dependence	Δ_{mass} depends on V_{ℓ_t} (tokenization and script coverage).	False gains if V_{ℓ_t} overcounts shared tokens; comparability can break.	Define V_{ℓ_t} by LID-labeled corpora; show Δ_{lid} alongside Δ_{mass}.
L2 Mid-resource bias	FOXP2 amplifies *existing* parametric memory; cannot create missing competence.	For low-resource languages, circuit may be weak; gains saturate early.	State scope explicitly; include "no-signal" diagnostic (flat spectrum / low lift).
L3 Semantic invariance	Defaultness gains meaningful only if content is preserved under the edit.	Edits can shift style/verbosity or factuality even when language matches.	Report task-suite ΔS; include QA exact match, summarization overlap checks.
L4 Safety interaction	Language steering can interact with safety filters not equally calibrated across languages.	Potential misuse: rephrasing unsafe content; bypassing refusal heuristics.	Add multilingual safety regressions; report harmful-content probes post-edit.
L5 Distribution shift	Matched-meaning prompts approximate "language-only" changes; real prompts are noisier.	Domain shift can change which features fire; long contexts dilute the signal.	Evaluate out-of-grid prompts; vary prompt length; report robustness across policies.
L6 Not a drop-in replacement	FOXP2 is an inference-time control knob, not a substitute for language-centric training.	Over-reading as "monolingual training unnecessary" risks under-investing in low-resource.	Make the non-claim explicit; recommend pairing with continued training where needed.
Roadmap (high-level, testable directions)			
FW Next steps	Standardize defaultness measurement + mechanistic edits as an auditable multilingual control primitive.	Must persist across models, tokenizers, and domains.	(1) Extend to more languages; (2) cross-model $\mathcal{N}/\mathcal{S}$ transfer; (3) publish stability protocol + safety pack.

Table 17: **Discussion & limitations at a glance (Neural FOXP2).** A scannable guide to interpret the three-stage pipeline (support $\mathcal{N}_{\ell_t}$, directions $\mathcal{S}_{\ell_t}^{(\ell)}$, window $\mathcal{W}$): what each component certifies, what can break, and which checks mitigate each risk.

faultness behaves like a **sparse, low-rank control circuit** in feature space: a small support $\mathcal{N}_{\ell_t}$ and a compact steering subspace $\mathcal{S}_{\ell_t}^{(\ell)}$ explain most consistent English$\rightarrow$target shifts (cf. sharp spectral concentration in Stage II), consistent with the broader view that high-level behaviors can be mediated by **few effective directions** even when representations are superposed (Elhage et al., 2022).

- **Interventional efficacy with bounded regression.** A **signed, sparse edit** constrained to $(\mathcal{N}_{\ell_t}, \mathcal{S}_{\ell_t}^{(\ell)}, \mathcal{W})$ produces a large early-step defaultness gain $(\Delta_{\mathrm{mass}}, \Delta_{\mathrm{lid}}, \Delta_{\mathrm{gain}})$ while keeping **task utility** essentially unchanged (Table 1), aligning with evidence that inference-time activation steering can deliver targeted behavior changes without full fine-tuning when interventions are structurally aligned (Turner et al., 2024)

- **Dictionary basis as a clean language coordinate system.** Language identity is more cleanly separable in a **dictionary / feature basis** than in raw channels: this yields an auditable **Language Feature Dictionary** and a stable intervention interface, matching the motivation for using sparse feature decompositions as a more interpretable unit than polysemantic neurons (Bricken et al., 2023).

- **Ablations behave mechanistically.** Random-feature edits fail, out-of-window edits degrade stability, and $\mathcal{N}$-only vs. $\mathcal{S}$-only edits expose complementary roles of **sparsity** and **geometry** (Table 1); the pattern is consistent with a **specific control set** rather than generic perturbation.

- **Language channels are partly separable.** Leakage checks (Hi$\rightarrow$Es) remain small under Hindi steering (and vice versa), suggesting that language-defaultness circuits exhibit **partial factorization** across languages rather

than a single undifferentiated "non-English" mode, consistent with mechanistic evidence that multilingual behavior can concentrate in **small, language-linked subsets** (Tang et al., 2024; Kojima et al., 2024b).

G.2 Why It Works

Localization → geometry: from *where* to *how*. Stage I answers *where* language control lives by returning a sparse support $\mathcal{N}_{\ell_t}$ of dictionary features that are simultaneously language-selective and causally linked to early-step defaultness. Stage II then answers *how* to move: it estimates the **dominant language-change directions** *restricted to the same coordinates we will edit*, by factorizing the paired language-shift matrix $\Delta Z^{(\ell,\ell_t)}$ into a small number of right-singular directions. This restriction is crucial: without it, SVD would mix language variation with irrelevant semantic/task axes that also differ across prompts (even when matched), weakening both stability and edit precision.

Low-rank interpretation (why layerwise SVD yields consistent directions). Let $\Delta\tilde{z}_k^{(\ell,\ell_t)} = \Pi_{\mathcal{N}_{\ell_t}^{(\ell)}} \Delta z_k^{(\ell,\ell_t)}$ be the localized feature-space shift for matched meaning unit k. If language defaultness is mediated by a low-dimensional control mechanism, then across diverse matched meanings the shifts share a common subspace:

$$\Delta\tilde{z}_k^{(\ell,\ell_t)} \approx \sum_{i=1}^{r_\ell} \alpha_{k,i}^{(\ell)} v_i^{(\ell,\ell_t)} + \varepsilon_k^{(\ell)}, \quad r_\ell \ll |\mathcal{N}_{\ell_t}^{(\ell)}|.$$

Stacking over k yields $\Delta Z^{(\ell,\ell_t)} \approx A^{(\ell)} V^{(\ell)\top} + E^{(\ell)}$, so the leading right singular vectors recover a maximally shared subspace of shifts. Intuitively, SVD finds the directions that are **most consistent across matched meanings**, precisely what a language-selection circuit should induce: it changes *language realization* while leaving *semantic content* (and therefore many semantic axes) invariant. This is the representational analog of isolating a **shared control mode** under varying inputs, consistent with the broader mechanistic view that superposed representations can still admit **few dominant functional directions** (Elhage et al., 2022).

Signed edit interpretation (push + suppress). Once $\mathcal{S}_{\ell_t}^{(\ell)}$ is identified, the signed sparse edit operationalizes two complementary effects: (i) **align** activations with the English→target shift directions (positive push), and (ii) **reduce** the competing English-default attractor by subtracting components that oppose the target shift (negative suppression), implemented within the same localized coordinate support. This matches the activation-engineering perspective: behavior can be steered by additive updates that align the hidden state with a direction associated with the desired behavior while avoiding broad distribution shift (Turner et al., 2024). The difference is that FOXP2 makes the direction *language-specific* (from paired multilingual shifts) and makes the support *sparse and auditable* (from dictionary localization) (Bricken et al., 2023).

Why early-step mass is the right lever (defaultness is a first-token phenomenon). Language defaultness is a **prior choice** that manifests at the beginning of generation: once the model emits a few target-language tokens, autoregressive conditioning makes continuing in that language easier, and later tokens are increasingly dominated by content and local syntax. Therefore the earliest steps ($t \in \{1, 2, 3\}$) provide the cleanest diagnostic for the models **language-selection prior**. In contrast, later decoding confounds language choice with task difficulty, lexical constraints, and self-induced distribution shift from previous tokens. FOXP2 explicitly targets this prior by optimizing shifts that increase $\Delta M^{\ell_t}(x, t)$ at the start, and then checks that downstream task utility does not regress (Table 1).

Mechanism sketch (3–5 lines).

> **Stage I:** find a sparse control support $\mathcal{N}_{\ell_t}$ in a dictionary basis; **Stage II:** recover low-rank shift directions $\mathcal{S}_{\ell_t}^{(\ell)}$ and a stable band $\mathcal{W}$; **Stage III:** apply a signed sparse edit on $(\mathcal{N}_{\ell_t}, \mathcal{S}_{\ell_t}^{(\ell)}, \mathcal{W})$ to increase early-step ΔM while bounding task regressions.

G.3 What the Ablations Mean

Table 1 is not only a performance summary; it is **mechanistic evidence**. The ablation patterns explain *why* the full method works.

Interpreting rows as causal tests ("because" statements).

- **Random features fail because** defaultness gain is not a generic property of sparse perturbation; it requires the **localized language**

support $\mathcal{N}_{\ell_t}$ (a specialized circuit rather than arbitrary sparsity).

- **Edits outside $\mathcal{W}$ weaken and destabilize because** language-change directions are not uniformly expressed across depth; there exists a contiguous band where the shift subspace is simultaneously **strong** (spectral mass / gain) and **reproducible** (bootstrap stability).
- **$\mathcal{N}$-only improves but saturates because** sparsity finds *where* to intervene but does not uniquely determine *how* to move; without $\mathcal{S}$, edits partially align with language-change geometry and partially spill into nuisance axes.
- **$\mathcal{S}$-only improves but leaks because** geometry alone identifies a direction but, when applied densely, it activates unrelated features and increases cross-language interference; sparsity is the precision instrument that keeps the edit on the intended control set.
- **Leakage stays small under FOXP2 because** the method edits language-specific supports and directions, suggesting partially separable circuits across Hindi and Spanish rather than a single "non-English" toggle (Tang et al., 2024; Kojima et al., 2024b).

A scannable summary of what each component of Neural FOXP2 certifies, what can break, and which checks mitigate each risk is provided in Table 17 in the appendix.

G.4 Relationship to Multilingual Training and Language-Specific Models

Neural FOXP2 is best interpreted as **parametric-memory reweighting**: it increases the probability that a multilingual model **accesses** and **executes** language-relevant structure that is already encoded, rather than learning that structure from scratch. This positioning matters because modern large backbones (e.g., LLaMA-family) are trained on broad web-scale mixtures dominated by CommonCrawl-like sources, so high-resource and mid-resource languages often exist in the training distribution but are **underweighted** relative to English. In parallel, language-specific models (e.g., Hindi-focused or Spanish-focused models) frequently draw from overlapping web data but differ in **sampling**, **curation**, and **weighting**, which can strengthen defaultness and stylistic naturalness in the target language. FOXP2 targets precisely this gap: it does not claim to add new linguistic competence; it claims to **amplify the already-present submanifold** and make it the models **default** under weak prompting.

What we did not claim (and where FOXP2 is most appropriate). FOXP2 is not an argument against monolingual or target-centric training. For **low-resource languages**, the base model may not contain sufficient lexical coverage, grammar robustness, or domain adaptation; steering cannot conjure missing competence. The method is therefore most plausibly helpful for **mid-to-high resource** languages (e.g., Hindi, Spanish, Portuguese, Russian) where substantial capacity is already present but **suppressed by defaultness**. In such cases, FOXP2 offers a complementary alternative to global fine-tuning: a mechanistic procedure to reweight language access while preserving shared multilingual and safety capabilities.

G.5 Practical Usage: Knobs, Operating Points, and Guardrails

The control knob. FOXP2 exposes an explicit edit intensity (or per-layer schedule) through the magnitude of $\delta z_{\ell_t}^{(\ell)}$ applied within $\ell \in \mathcal{W}$. Operationally, users can treat the overall scale (and optional layerwise weights) as a **single knob** that trades off defaultness gain against any downstream regression, analogous in spirit to controlling strength in activation-addition style interventions (Turner et al., 2024).

Safe operating point. A principled operating point is the solution to a constrained selection rule: maximize early-step defaultness gain subject to utility and stability constraints,

$$\begin{aligned} &\max \Delta_{\text{gain}} \\ &\text{s.t.} \quad \Delta S \geq -\tau_{\text{util}}, \\ &\qquad \text{Boot. Stab} \geq \tau_{\text{stab}}, \\ &\qquad \text{Leakage} \leq \tau_{\text{leak}}. \end{aligned}$$

where Δ_{gain} is computed over neutral prompts, and ΔS summarizes downstream task utility (Table 1). This makes deployment decisions legible: the chosen edit strength is not aesthetic, but **constraint-certified** by regression checks.

When to use FOXP2 vs. prompting vs. fine-tuning (decision guide).

- Use **prompting** when language instruction is explicit and reliable, and you do not need persistent defaultness under weak prompts.
- Use **Neural FOXP2** when you need **default-language behavior** (first-token prior) without training a new checkpoint, and you can enforce regression guardrails.
- Use **fine-tuning** when target-language competence is missing (low-resource, domain shift, or systematic errors), i.e., when the issue is **knowledge/skill acquisition** rather than **language selection**.

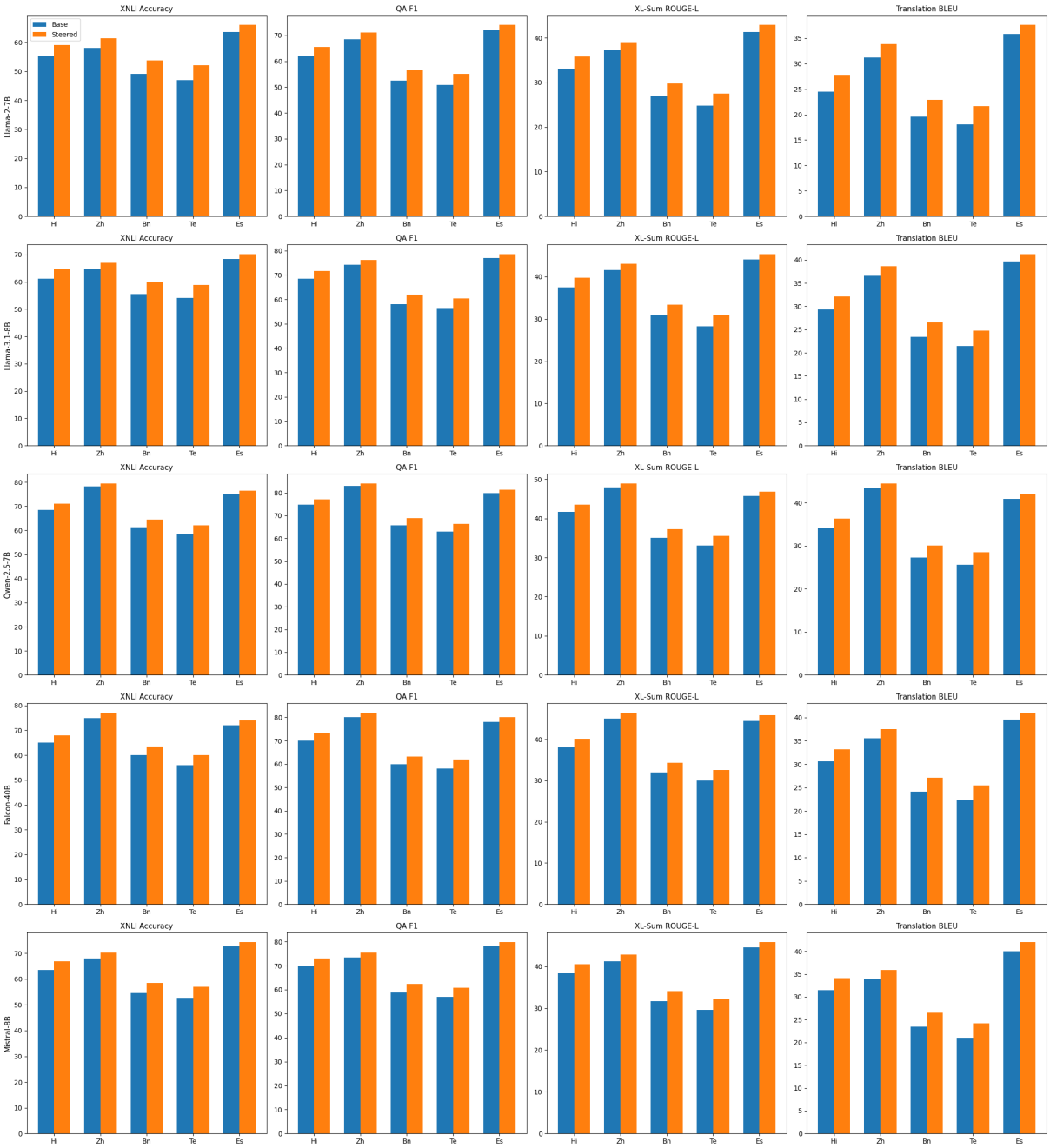

Figure 16: Comparative multilingual evaluation between the baseline and neuron-steered models across XNLI Accuracy, QA F1, XL-Sum ROUGE-L, and Translation BLEU. Results are reported for five representative languages: Hindi (Hi), Chinese (Zh), Bengali (Bn), Telugu (Te), and Spanish (Es). Across all evaluated models, steering language-specific features consistently improves multilingual understanding, summarization, and translation performance.