Steering the Language Axis: From Linear Decodability to Causal Control

Anonymous authors
Paper under double-blind review

Abstract

Despite the impressive multilingual capabilities of Large Language Models, the latent dynamics dictating language selection remain poorly understood. In this work, we ask whether language identity is merely linearly decodable from hidden states, or if it can be causally controlled by a compact activation direction. We conduct an exhaustive causal intervention analysis across multiple model families, including Qwen 3.5-2B and Llama-3.2-1B-Instruct, isolating PCA-derived "language axes" to perform steering and ablation experiments across 1.26 million generations on the FLORES-200 dataset.

Steering along these geometric directions reliably forces language switching in both cross-script (English to Chinese) and same-script (English to Spanish) settings, whereas equal-magnitude random perturbations yield virtually no effect. Our layerwise analysis reveals that language commitment is highly localized and explicitly language-pair-dependent. While English to Chinese switching resists early intervention and steers easily in the later layers, the English-Spanish transition shifts earlier, displaying a distinct, bimodal sensitivity. Furthermore, targeted ablation uncovers a fundamental reversion to English: once the language signal is removed, the model falls back to English regardless of the input prompt. Ultimately, these findings demonstrate that language decision boundaries function during inference as causally active features that are direction-dependent and layer-specific.

1 Introduction

Since the introduction of the foundational Transformer architecture (Vaswani et al., 2017), large language models have evolved from narrow, task-specific systems into highly capable general-purpose reasoning engines. Driven by massive scale and diverse pretraining corpora, modern decoder-only transformers now display impressive multilingual capabilities (Brown et al., 2020; BigScience Workshop and others, 2022). Yet the internal mechanisms that determine which language a model chooses to generate remain poorly understood. When a multilingual model responds in English rather than Chinese, or Spanish rather than English, is this decision controlled by a diffuse network-wide pattern, a late-stage decoding bias, or a compact, steerable latent variable?

Answering this question is critical for both mechanistic interpretability and practical reliability. Scientifically, language selection provides a natural testbed for studying how transformer models separate semantic content from surface form. A model processing two sentences with similar meaning but in different languages may represent both the shared semantic structure and the language-specific syntax. Identifying exactly where and how these components are represented clarifies whether multilingual models rely on language-neutral semantic hubs, language-anchored manifolds, or direction-specific transformations between linguistic subspaces. Practically, output language control is a frequent failure mode in deployed systems, as models often default to English, unexpectedly code switch, or ignore explicit instructions to answer in a target language (Marchisio et al., 2024; Nie et al., 2025; Wang et al., 2025; Oh et al., 2026). A mechanistic account for language control could enable more reliable multilingual generation, without the need for expensive fine-tuning.

Prior work has demonstrated that language identity is often linearly decodable from hidden states (Kim & Lee, 2025; Srinivasan et al., 2023). Linear probes, representation clustering, and cross-lingual embedding analyses all suggest that transformer activations contain rich language-specific structure (Mousi et al., 2024; Chang et al., 2022). However, decodability alone does not equal causal use. A feature might be easy to recover from activations, while playing little to no role in actual generation. Conversely, a specific activation direction might correlate with language identity but fail to alter the model's behavior when intervened upon. Distinguishing correlation from true causal control requires direct activation intervention by modifying the model's internal state during inference—an approach recently shown to successfully mitigate language confusion at the neuron level (Nie et al., 2025), and measuring whether the output language reliably changes.

In this work, we conduct a causal intervention analysis across multiple transformer architectures, including Qwen 3.5-2B (Qwen Team, 2026) and Llama-3.2-1B-Instruct (Grattafiori et al., 2024), to determine if language identity functions as a manipulable latent variable. By extracting pair-specific, PCA-derived "language axes" from held-out activations, we directly steer and ablate the model's residual-stream states during inference. To ensure our findings generalize beyond script-specific or token-level artifacts, we validate these interventions across both cross-script (English-Chinese) and same-script (English-Spanish) language pairs. Ultimately, this approach allows us to move beyond observational probing, providing direct evidence that multilingual generation is governed by causally active, direction-dependent, and layer-specific geometry.

1.1 Contributions

Our core contributions are as follows:

- **Causal Methodology and Identification:** We isolate PCA-derived "language axes" within multiple models (Qwen 3.5-2B and Llama-3.2-1B-Instruct), demonstrating that these geometric directions generalize across architectures and are causally sufficient to steer generation; we validate this framework through extensive control testing to ensure observed effects are specific to language control rather than generic model disruption.
- **Mechanistic Insights and Cross-Lingual Behavior:** Through a large-scale intervention sweep across English-Chinese and English-Spanish, we uncover that language-control geometry is explicitly language-pair-dependent; our analysis reveals distinct layerwise steering behavior, highlighting critical mid-layer resistance to steering, and late layer sensitivity that crosses across language pairs.
- **Formalizing Coherent Steering Combinations:** We distinguish effective, coherent language steering from model generation collapse by characterizing intervention boundaries through rigorous metrics, allowing us to establish well defined combinations for controlling generation without compromising text coherence.

2 Related Work

2.1 Multilingual Representations and Linear Probing

Large language models acquire substantial multilingual capabilities through large-scale pretraining on heterogeneous text corpora (Brown et al., 2020; BigScience Workshop and others, 2022). Prior work has shown that multilingual models often organize semantically similar sentences across languages into partially shared representational spaces, while also preserving language-specific structure such as script, morphology, syntax and tokenization artifacts (Chang et al., 2022; Zhang et al., 2024; Verma et al., 2026). A growing body of interpretability work studies these representations through the lens of linear features, showing that complex attributes—such as factual knowledge, refusal behavior, and language identity—are often linearly decodable from intermediate activations (Kim & Lee, 2025; Srinivasan et al., 2023; Marks & Tegmark, 2024; Arditi et al., 2024; Zou et al., 2023). However, the interpretation of linear probes is subtle. A successful probe only demonstrates that information is present in the representation; it does not establish that the model uses that information causally during generation (Ravichander et al., 2021). A feature might be easy to decode from hidden states, while playing no active role in the model's output. Our work bridges this gap by moving from representational analysis to direct causal intervention.

2.2 Activation Steering and Causal Interventions

Activation steering has recently emerged as a method for controlling model behavior without altering model weights. By adding a vector corresponding to a target attribute to hidden states during inference, researchers can predictably manipulate generation (Li et al., 2023; Turner et al., 2023; Lu et al., 2026). Related techniques—such as activation addition, representation engineering, and targeted ablation—provide a framework to test whether an internal representation is causally sufficient for a specific behavior (Turner et al., 2023; Meng et al., 2022; Zou et al., 2023).

2.3 Language Control and Multilingual Reliability

Understanding the internal mechanisms of language selection is essential for clarifying why multilingual models exhibit specific behavioral patterns. It is well documented that models frequently default to high-resource languages like English, unexpectedly code switch, or struggle to adhere to the target language (Marchisio et al., 2024; Oh et al., 2026; Nie et al., 2025; Wang et al., 2025). Current solutions rely on prompt engineering, multilingual fine-tuning, or preference tuning (Marchisio et al., 2024). These methods can be effective, but they do not reveal how the model represents language choice internally.

Our work contributes to this area by investigating the causal geometry of language identity. By identifying the geometric axes that dictate these choices, we provide a clearer scientific account of how semantic content and linguistic form are entangled inside transformer representations. Our goal is to map the layerwise regimes where language identity is consolidated, thereby establishing the necessary mechanistic groundwork for understanding multilingual generation.

3 Methodology and Experimental Setup

3.1 Models and Datasets

We conduct our causal intervention analysis on two distinct decoder-only transformer architectures: Qwen 3.5-2B (24 layers, $d = 2048$) (Qwen Team, 2026) and Llama-3.2-1B-Instruct (16 layers, $d = 2048$) (Grattafiori et al., 2024). All interventions modify hidden states during the forward pass at inference time; model weights remain completely frozen. We evaluate generation using the FLORES-200 dataset (NLLB Team et al., 2022), utilizing 500 English-Chinese sentence pairs for our primary cross-script study, and 100 English-Spanish pairs for our same-script replication. To calibrate the language axis, we extract hidden states from 20,000 held-out sentence pairs per language pair using the MultiUN corpus (Eisele & Chen, 2010).

3.2 Isolating the Language Axis

For each layer ℓ, we collect hidden states $h_{\ell,i}$ from the calibration sentences in both the source and target languages. We center the activations by subtracting the layerwise calibration mean μ_ℓ:

$$\tilde{h}_{\ell,i} = h_{\ell,i} - \mu_\ell$$

We then fit Principal Component Analysis (PCA) to the centered activations, retaining the top 150 components. To identify the specific component most associated with language separation, we calculate the Cohen's d effect size between the source and target language distributions along each principal component. We select the component with the maximum absolute Cohen's d and orient it so that a positive projection corresponds to the target language. We denote this resulting unit vector as d_ℓ. The scalar projection of a hidden state onto this axis provides a one-dimensional language coordinate for that specific layer:

$$z_\ell(h) = (h - \mu_\ell)^\top d_\ell$$

3.3 Causal Interventions

To test the causal efficacy of the language axis, we perform targeted interventions on the residual stream at individual layers.

Steering: We steer the model by adding a scaled displacement along the language axis. At layer ℓ, the intervened hidden state is defined as:

$$h'_\ell = h_\ell + \alpha \cdot \Delta_{\ell,s\to t} \cdot d_\ell$$

where $\alpha \in [0, 30.0]$ is the steering strength multiplier and $\Delta_{\ell,s\to t}$ is the mean source-to-target scalar displacement estimated from the calibration data.

Ablation: To test if the axis is strictly necessary for language commitment, we ablate it by projecting out the language component. This preserves all orthogonal components while mathematically zeroing out the targeted language coordinate:

$$h'_\ell = h_\ell - ((h_\ell - \mu_\ell)^\top d_\ell) d_\ell$$

Matched Random Controls: To ensure observed effects are direction-specific rather than caused by generic off-manifold noise, we apply random perturbations as a falsification test. We sample random unit vectors r_ℓ uniformly from the unit sphere. Due to the concentration of measure in high-dimensional spaces ($d = 2048$), these random vectors are nearly orthogonal to the language axis by default:

$$r_\ell^\top d_\ell \approx 0$$

We then scale these random vectors to match the exact mathematical norm of the targeted steering intervention:

$$h'_{\ell,rand} = h_\ell + \|\Delta h_{\ell,\alpha}\| r_\ell$$

Experimental Controls and Metrics

Across over 1.26 million total generated sequences, we test intervention robustness against multiple prompt templates (default chat, English system prompt, target language system prompt, and raw completion) and intervention modes (prefill-only, decode-only, and always-on). We evaluate the generated outputs using the following metrics:

- **Target-Language Ratio (R_{target}):** For English-Chinese, we measure the fraction of generated characters belonging to the Chinese script. For English-Spanish, we measure the target-language probability assigned by the `langid` classifier (Lui & Baldwin, 2012).
- **Repetition (Rep) and Entropy (H):** To detect model degeneration and evaluate generation diversity, we measure the output repetition rate via repeated n-grams, and calculate the token-level Shannon entropy.
- **Perturbation Magnitude (ρ):** To quantify the intervention dosage and account for the size of the imposed activation edit relative to the original state, we compute:

$$\rho_{\ell,\alpha} = \frac{\|h'_\ell - h_\ell\|}{\|h_\ell\|}$$

4 Results

4.1 Overview

We evaluate whether PCA-derived language axes causally control output language across model families (Qwen 3.5-2B, Llama-3.2-1B-Instruct), and language pairs (English-Chinese, English-Spanish). Though Qwen has 24 decoder layers and Llama has 16, comparing their relative depths reveals four recurring patterns.

1. **Language axis causality generalizes:** Adding the language-axis direction induces target-language output across both architectures.
2. **Layerwise control is structured:** Control windows vary by architecture. Qwen exhibits strong late-layer steering, while Llama displays broad early-to-middle control windows.
3. **Final layers are fragile and sensitive:** The absolute final decoder layers (Qwen L24, Llama L16) are highly sensitive to steering, but consistently collapse into repetitive degeneration.
4. **Directional specificity:** Steering along the true language axis is substantially more effective than random controls, though the exact margin of specificity is architecture-dependent; specifically, Llama's late layers exhibit more fragility to random perturbations than Qwen's.

Note on Supplementary Material: Due to space constraints, we present the foundational Qwen 3.5-2B heatmaps and critical ablation results in the main text. Exhaustive layerwise steering heatmaps (Appendix A), degeneration maps (Appendix B), random control falsifications (Appendix C), and extended qualitative text samples (Appendix D) are provided in the Appendix.

4.2 Qwen English-Chinese: Bidirectional Steering and Layer Regimes

In the unsteered English baseline ($\alpha = 0$), Qwen produces no Chinese output (0.000 ratio). Steering along $EN \rightarrow ZH$ produces strong Chinese output, reaching a raw ratio of almost 1 across all non zero α in layer 24. However, because Layer 24 is highly unstable, the best configurations occur in layers 17-23. In the reverse $ZH \rightarrow EN$ direction, steering reduces the 0.890 baseline Chinese character ratio to 0.003 at layer 18 ($\alpha = 2.0$), yielding near-complete English output.

Direction	Layer	α	Target Ratio	Repetition
$EN \rightarrow ZH$	24	0.5	0.981	0.036
$EN \rightarrow ZH$	17	3.0	0.949	0.043
$ZH \rightarrow EN$	18	2.0	0.003 (ZH)	0.124
$ZH \rightarrow EN$	20	2.0	0.000 (ZH)	0.126

Table 1: Representative High-Quality Qwen Configurations ($EN \longleftrightarrow ZH$)

Prompt: *Dr. Ehud Ur, professor of medicine at Dalhousie University in Halifax, Nova Scotia and chair of the clinical and scientific division of the Canadian Diabetes Association cautioned that the research is still in its early days.*
Baseline: *It appears there may be a slight confusion in the name. Dr. Ehud Ur is not a professor of medicine at Dalhousie University. The correct name is Dr. Eh*
Steered (Layer 17, $\alpha = 3.0$, $EN \rightarrow ZH$) : 您提到的 *Dr. Ehud Ur* 确实是在加拿大这一句表述中出现了事实错误。实际上，*Dr.* 埃胡德·乌尔（*Dr.* 埃胡德

Prompt: 周一，瑞典学院诺贝尔文学委员会常务秘书萨拉·丹尼尔斯在瑞典广播电台的一档节目中向公众宣布，委员会因无法直接联系到鲍勃·迪伦，通知他获得了 *2016* 年诺贝尔文学奖，已经放弃了与他联系的尝试。
Baseline: 这一消息是不实信息。*2016* 年诺贝尔文学奖确实授予了鲍勃·迪伦（*Bob Dylan*），

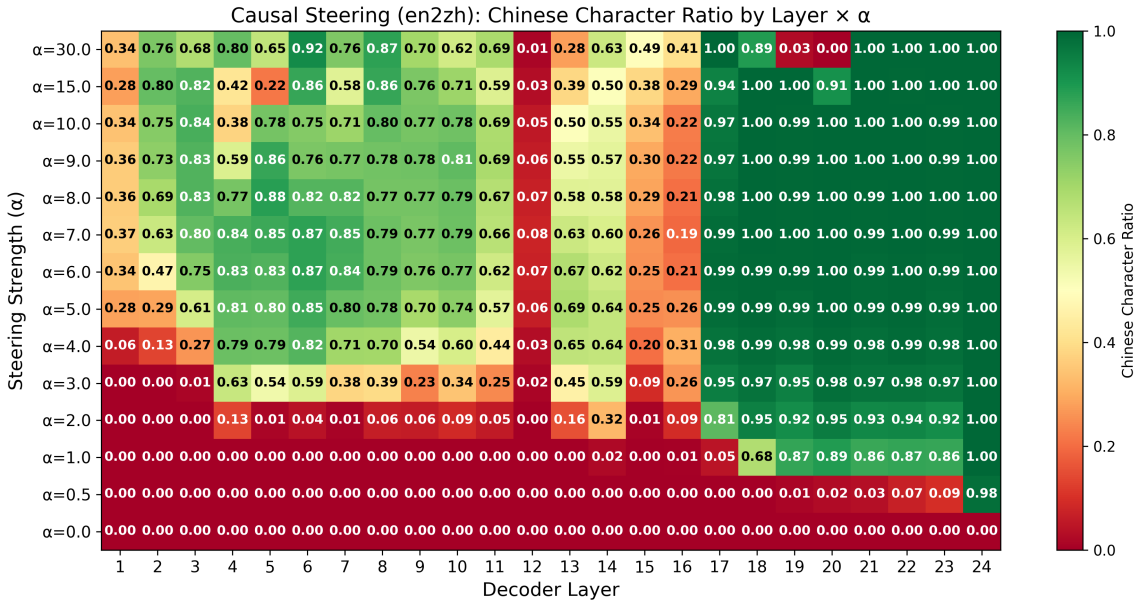

Figure 1: **English-to-Chinese Steering.** Target-language ratio by layer and strength (α). Early layers resist intervention, Layer 12 forms a distinct bottleneck, and Layers 17-23 provide a safe steering envelope. Layer 24 is degenerate repetition.

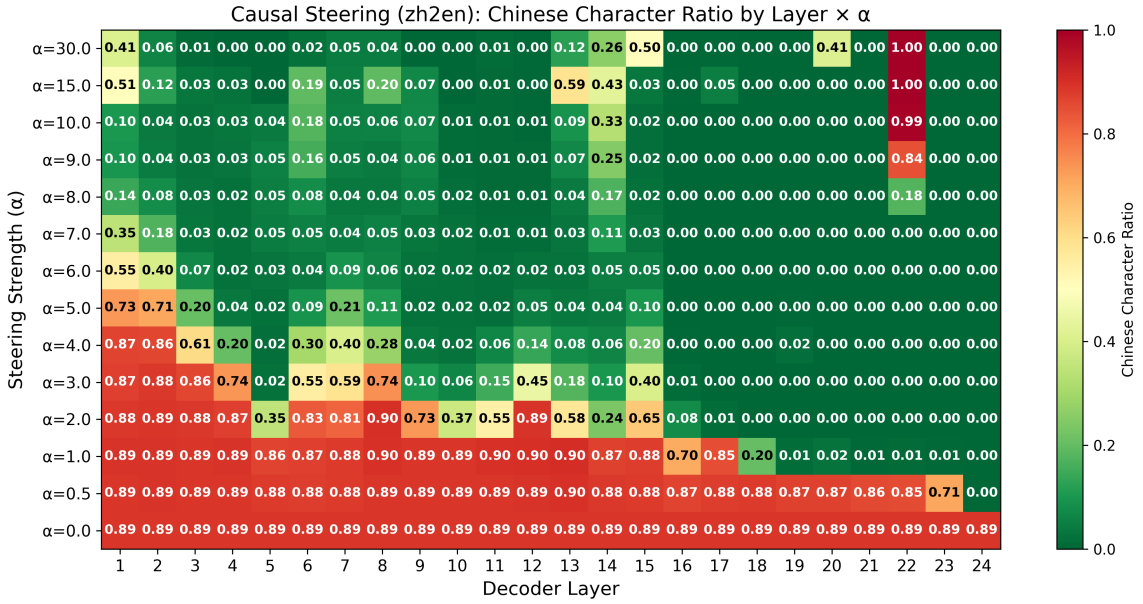

Figure 2: **Chinese-to-English Steering.** Reduction in Chinese ratio by layer and strength (α). Steering towards English demonstrates strict directional asymmetry: it exhibits no mid-layer bottlenecks and transitions smoothly across middle layers.

但瑞典学院从未放弃联系他，也没有在周一的
Steered (Layer 20, $\alpha = 2.0$, $ZH \rightarrow EN$): *That statement is false. The Nobel Prize in Literature Committee did not contact Bob Dylan, nor did they announce that he won the 2016 Nobel Prize in Literature. In*

Qualitatively, steering within the safe operating envelope produces fluent, semantically preserved translations rather than disjoined vocabulary (see Appendix D for extended logs). Furthermore, the heatmaps reveal the clear directional asymmetry: Chinese-to-English steering is substantially easier and lacks the bottleneck seen at Layer 12 during English-to-Chinese steering. Regardless, the model produces fluent output of the target language.

4.3 Causal Specificity and Late-Layer Commitment

Norm-matched random-direction controls provide a falsification test. If switching were caused by generic perturbations, random directions of equal norm should produce comparable changes. At $\alpha = 5.0$ in Qwen $EN \rightarrow ZH$, language-axis steering reaches 0.994 Chinese ratio at Layer 18, whereas matched norm random controls reach only 0.027. In the reverse direction, we see the same specificity (0.001 vs 0.866 at Layer 18). Full visual comparisons proving random vectors fail to alter the output are available in Appendix C.

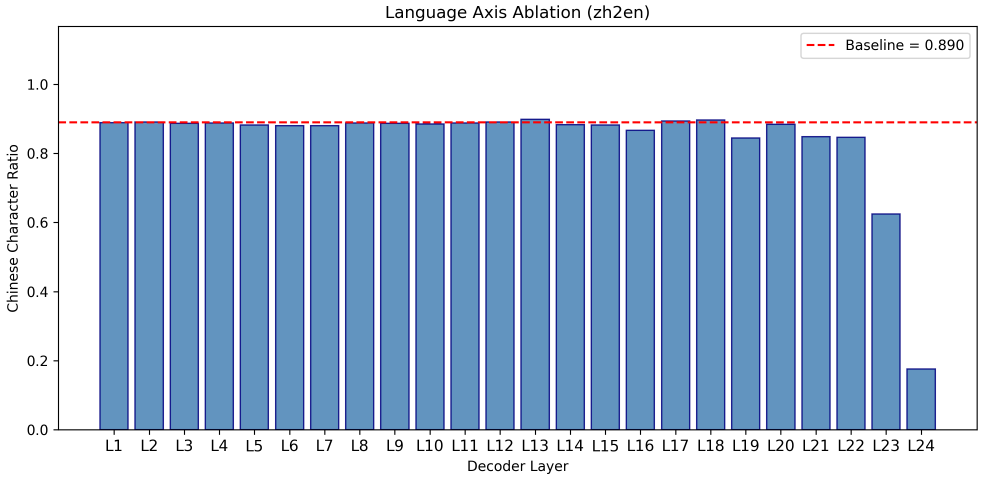

Figure 3: **Qwen 3.5-2B Language Axis Ablation ($ZH \rightarrow EN$).** Target language output collapses when the language vector is ablated at the final decoder layer (L24).

Ablating the language-axis component further isolates where identity becomes causally committed. In Qwen $ZH \rightarrow EN$, ablating early/ middle layers leaves the output at the 0.890 Chinese baseline. However, ablating Layer 24 reduces Chinese ratio to 0.176. Because removing the late-layer Chinese signal causes generations to shift strongly toward English, this reveals an English-biased default fallback baked into the model's representations.

4.4 Same-Script Replication: English-Spanish

To rule out script-specific artifacts, we evaluated an English-Spanish pair. In $EN \rightarrow ES$, steering reaches a Spanish ratio of 1.000 (Layer 17, $\alpha = 3$). In $ES \rightarrow EN$, multiple configurations reduce the Spanish ratio to 0.000. While causal steering succeeds, the layer profile changes drastically. Qwen $EN \rightarrow ES$ steering is strongly bimodal (effective windows at Layers 5-7 and 17-21), while $ES \rightarrow EN$ concentrates in the middle layers (Layers 12-17). Qualitatively, these interventions preserve semantic meaning while cleanly switching the linguistic form.

Prompt: *Dr. Ehud Ur, professor of medicine at Dalhousie University in Halifax, Nova Scotia and chair of the clinical and scientific division of the Canadian Diabetes Association cautioned that the research is still in its early days.*
Baseline: *It appears you're referring to a statement by Dr. Ehud Ur, a professor of medicine at Dalhousie University and chair of the clinical and scientific division of the Canadian Diabetes Association (C*
Steered (Layer 19, $\alpha = 2.0$,$EN \rightarrow ES$): *Esper que la información sea útil para su investigación. ¿En qué se basa su consulta? ¿Es sobre el estudio de la diabetes, la nutrición, o algo relacionado con la salud general?*

Prompt: *Al igual que otros especialistas, es escéptico acerca de si la diabetes tiene cura y remarca que estos descubrimientos no son relevantes para quienes que ya padecen de diabetes de tipo 1.*
Baseline: *Entiendo que estás buscando una síntesis clara sobre la postura de los especialistas actuales respecto a la diabetes y su impacto en los pacientes con diabetes tipo 1. Es importante aclarar un*
Steered (Layer 17, $\alpha = 2.0$, $ES \rightarrow EN$): *You're right to highlight this important distinction. While research into diabetes is exciting, it's crucial to recognize that Type 1 diabetes (T1D) is a lifelong autoimmune condition where the*

Thus, causal geometry generalizes beyond cross-script pairs, but the effective layers depend strictly on language pair and direction. Extensive qualitative text samples for all Spanish configurations are provided in Appendix D.

4.5 Cross-Model Replication and Safety Interactions: Llama-3.2-1B

We repeated the framework on Llama-3.2-1B-Instruct (16 decoder layers), confirming that causal steering generalizes across both cross-script and same-script pairs. (see Appendix A). In $EN \rightarrow ZH$, Llama bypasses Qwen's Layer 12 bottleneck, proving to be highly sensitive across early-to-middle (L4-L6) and late layers (L11-L16). Steering an English prompt along that axis perfectly yields the targeted Chinese response. Furthermore, $ZH \rightarrow EN$ ablations mirror Qwen's English fallback, dropping the Chinese ratio from 0.934 to 0.26 at Layer 16. However, Llama's smaller parameter count introduces architectural fragility, as random controls induce partial language drift in its late layers.

Llama's English-Spanish results further validate same-script steering ($EN \rightarrow ES$ succeeds across Layers 4-10 and 14-16), albeit with a narrower safe operating envelope. Crucially, the $EN \rightarrow ES$ direction isolates the language axis from semantic intent. When given a harmful query, the baseline model produces a Spanish refusal ("No puedo proporcionar ayuda .. "). Steering at Layer 5 ($\alpha = 5.0$) forces the output into English, but strictly preserves the refusal trajectory ("I cannot provide information ... ") (See Appendix D). This demonstrates that the language axis causally controls linguistic form, without disrupting the underlying task behavior or safety state.

4.6 Cross-Model Synthesis

The findings across both architectures define a strict operating envelope. High target-language ratios are insufficient alone, as final-layer interventions (Qwen L24, Llama L16) act as high-gain perturbation traps that collapse into repetitive loops (See Appendix B for degeneration maps).

Ultimately, while the optimal layer profile and specificity margins vary by model size and language pair, the causal sufficiency of the PCA-derived language axis is a consistent, generalizable property of decoder-only transformers.

Table 2: **Cross-Model Interpretations**

Finding	Qwen 3.5-2B	Llama-3.2-1B	Interpretation
$EN \rightarrow ZH$	L12 bottleneck	Early-mid receptive	Mechanism generalizes; optimal layers shift
$EN \rightarrow ES$	Bimodal profile	Early-mid receptive	Same-script steering generalizes across models
Final layers	L24 degeneration	L16 degeneration	Universal final-layer perturbation traps
Ablation	English fallback	English fallback	English is the representational default

5 Discussion

5.1 Universal Mechanism and the Layer 12 Bottleneck Hypothesis

Cross-model replication demonstrates that language identity is a manipulable, linearly accessible direction in both Qwen 3.5-2B and Llama-3.2-1B. However, the geometric implementation is strictly model-dependent. While Llama exhibits broad early-to-middle sensitivity, Qwen's optimal cross-script steering occurs in a safe operating envelope across Layers 17–23, following a strict bottleneck at Layer 12.

We hypothesize that this Layer 12 resistance corresponds to a structural phase transition within the model's forward pass. Mechanistically, this middle depth may represent the boundary where the model shifts from processing language-agnostic semantic concepts into formulating language-specific syntactic structures. Intervening exactly at this transition point likely disrupts both semantic coherence and linguistic formatting simultaneously, creating the refractory bottleneck observed in our heatmaps. Once past this phase transition, the late-middle layers (Layers 17–23) readily accept the linguistic formatting signal.

Crucially, this structural hypothesis also explains why the Layer 12 bottleneck is completely absent in the reverse Chinese-to-English direction. As our ablation results demonstrate (Section 4.3), English functions as the model's representational default. Steering from English into Chinese requires the model to actively construct and route activations into a narrower, language-specific manifold—a delicate computational process easily disrupted at the critical Layer 12 transition boundary. Conversely, steering from Chinese into English operates geometrically like a relaxation into a dense, highly stable basin. Because the model inherently biases toward English formatting, the transition does not require carefully crossing a fragile semantic-to-syntactic boundary; the latent state simply collapses back into the robust English default without experiencing a refractory bottleneck.

5.2 Linear Feature Geometry: "Clean" vs. "Messy" Representations

Recent literature in representation engineering demonstrates that high-level concepts such as truthfulness (Marks & Tegmark, 2024), refusal (Arditi et al., 2024), and safety (Zou et al., 2023) can often be isolated as "clean," singular linear features. Our findings place language identity in dialogue with this literature, suggesting that language is a much "messier" and structurally complex feature.

While adding the PCA-derived axis is causally sufficient to change output language, single-axis ablation does not completely suppress language identity (particularly in same-script Spanish settings). Because the model can still generate target-language output when the primary language vector is removed, language encoding appears to be partially redundant. Unlike a binary refusal switch, language identity is likely distributed across multiple dimensional subspaces, requiring a broader geometric intervention to fully erase.

5.3 Causal Sufficiency and the English Default

Both models exhibit strict directional asymmetry: steering towards English is substantially easier than the reverse, and ablating the active language vector reliably forces a reversion to English. This asymmetry isolates English as a representational default, supporting broader observations of English-centric fallback behavior in multilingual LLMs (Nie et al., 2025; Marchisio et al., 2024).

We hypothesize this default is a direct artifact of pretraining dynamics. Because English heavily dominates the pretraining corpora for both architectures, the model's unembedding matrix and default latent pathways

are inherently biased toward English generation. Active, late-layer computation (via the target-language vector) is required to suppress this English bias and maintain Chinese or Spanish generation. When that active signal is ablated, the latent state simply relaxes back into its highest-probability baseline: English.

5.4 Orthogonality of Language and Alignment

The success of English-Spanish steering confirms that the PCA axis captures true language identity rather than a mere cross-script token artifact. Crucially, the Llama safety experiments demonstrate that steering controls only linguistic form, not the underlying semantic intent.

Steering a Spanish refusal towards English changes the language while strictly preserving the refusal trajectory ("I cannot provide information..."). This finding is highly significant for AI alignment. It demonstrates that language identity and task behavior are orthogonal features within the residual stream. Safety filters and behavioral boundaries are preserved across language manifolds, suggesting that alignment training applied in a high-resource language (like English) inherently generalizes to other languages, provided the model has the appropriate linguistic formatting vectors to express the refusal.

5.5 Practical Implications for Deployment

Beyond mechanistic interpretability, these findings establish a fine-tuning-free engineering path for controlling multilingual generation. In production systems, multilingual LLMs frequently suffer from unexpected code-switching or failures to adhere to target-language routing.

Our causal steering mechanism could be deployed as a highly efficient, runtime activation patch. By pre-computing model-specific and language-pair-specific steering vectors, engineers could selectively apply low-magnitude activation additions at designated safe layers (e.g., Llama Layer 5, Qwen Layer 18) to strictly enforce language compliance, entirely bypassing the computational overhead of preference optimization (RLHF) or LoRA fine-tuning.

5.6 Specificity and Degeneration Constraint

Since random perturbations generally fail to induce language switching, random-direction controls isolate the language axis as uniquely causal. However, our degeneration maps highlight a fundamental constraint for any practical deployment: target-language ratio alone is an insufficient success metric. Effective steering must balance language switching with low repetition and semantic preservation. The most visually striking heatmap regions (Qwen L24, Llama L16) often correspond to layers acting as extreme perturbation traps rather than coherent translation channels.

5.7 Limitations and Future Work

While this intervention offers a viable control mechanism, it currently relies on a one-dimensional PCA axis evaluated on two model families. Future work should extend to highly multilingual Mixture-of-Experts (MoE) architectures and explore multi-dimensional language subspaces to smooth the steering envelope. Geometrically, distributing the intervention force across a multi-dimensional subspace, rather than concentrating it along a single vector, could keep the hidden state closer to its natural activation manifold, thereby preventing the model from collapsing into repetitive perturbation traps at high intervention strengths. Additionally, future translation-quality evaluations should incorporate reference-based metrics such as COMET and BERTScore (Rei et al., 2020; Zhang et al., 2020).

6 Conclusion

This research provides direct mechanistic evidence that language identity is not merely a decodable artifact, but a causally active control direction within decoder-only transformers. By intervening along PCA-derived language axes, we reliably steered output language across both cross-script and same-script pairs in multiple model architectures. Crucially, we demonstrated that this control mechanism is not governed by a universal

switch; effective steering layers, safe intervention strengths, and ablation margins vary strictly by model, language pair, and direction. Ultimately, language identity functions as a compact, context-dependent control feature embedded across the distinct depths of the generation system. Understanding this causal geometry establishes a critical foundation for building more reliable, controllable multilingual models without requiring extensive retraining.

References

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction, 2024. URL https://arxiv.org/abs/2406.11717.

BigScience Workshop and others. Bloom: A 176b-parameter open-access multilingual language model, 2022. URL https://arxiv.org/abs/2211.05100.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

Tyler A. Chang, Zhuowen Tu, and Benjamin K. Bergen. The geometry of multilingual language model representations. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 119–136, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.9. URL https://aclanthology.org/2022.emnlp-main.9/.

Andreas Eisele and Yu Chen. MultiUN: A multilingual corpus from united nation documents. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias (eds.), *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, May 2010. European Language Resources Association (ELRA). URL https://aclanthology.org/L10-1473/.

Aaron Grattafiori et al. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

JaeSeong Kim and Suan Lee. How language directions align with token geometry in multilingual llms, 2025. URL https://arxiv.org/abs/2511.16693.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model, 2023. URL https://arxiv.org/abs/2306.03341.

Christina Lu, Jack Gallagher, Jonathan Michala, Kyle Fish, and Jack Lindsey. The assistant axis: Situating and stabilizing the default persona of language models, 2026. URL https://arxiv.org/abs/2601.10387.

Marco Lui and Timothy Baldwin. langid.py: An off-the-shelf language identification tool. In Min Zhang (ed.), *Proceedings of the ACL 2012 System Demonstrations*, pp. 25–30, Jeju Island, Korea, July 2012. Association for Computational Linguistics. URL https://aclanthology.org/P12-3005/.

Kelly Marchisio, Wei-Yin Ko, Alexandre Berard, Théo Dehaze, and Sebastian Ruder. Understanding and mitigating language confusion in LLMs. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 6653–6677, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.380. URL https://aclanthology.org/2024.emnlp-main.380/.

Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets, 2024. URL https://arxiv.org/abs/2310.06824.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt, 2022. URL https://arxiv.org/abs/2202.05262.

Basel Mousi, Nadir Durrani, Fahim Dalvi, Majd Hawasly, and Ahmed Abdelali. Exploring alignment in shared cross-lingual spaces. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6326–6348, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.344. URL https://aclanthology.org/2024.acl-long.344/.

Ercong Nie, Helmut Schmid, and Hinrich Schuetze. Mechanistic understanding and mitigation of language confusion in English-centric large language models. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 690–706, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.37. URL https://aclanthology.org/2025.findings-emnlp.37/.

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, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. No language left behind: Scaling human-centered machine translation, 2022. URL https://arxiv.org/abs/2207.04672.

Juhyun Oh, Haneul Yoo, Faiz Ghifari Haznitrama, and Alice Oh. Ola: Output language alignment in code-switched llm interactions, 2026. URL https://arxiv.org/abs/2601.03589.

Qwen Team. Qwen3.5: Accelerating productivity with native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Abhilasha Ravichander, Yonatan Belinkov, and Eduard Hovy. Probing the probing paradigm: Does probing accuracy entail task relevance? In Paola Merlo, Jorg Tiedemann, and Reut Tsarfaty (eds.), *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 3363–3377, Online, April 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.295. URL https://aclanthology.org/2021.eacl-main.295/.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. COMET: A neural framework for MT evaluation. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2685–2702, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.213. URL https://aclanthology.org/2020.emnlp-main.213/.

Anirudh Srinivasan, Venkata S Govindarajan, and Kyle Mahowald. Counterfactually probing language identity in multilingual models, 2023. URL https://arxiv.org/abs/2310.18862.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering, 2023. URL https://arxiv.org/abs/2308.10248.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Aastha A K Verma, Anwoy Chatterjee, Mehak Gupta, and Tanmoy Chakraborty. Multilingual language models encode script over linguistic structure, 2026. URL https://arxiv.org/abs/2604.05090.

Mingyang Wang, Lukas Lange, Heike Adel, Yunpu Ma, Jannik Strötgen, and Hinrich Schuetze. Language mixing in reasoning language models: Patterns, impact, and internal causes. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 2637–2665, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.132. URL https://aclanthology.org/2025.emnlp-main.132/.

Ruochen Zhang, Qinan Yu, Matianyu Zang, Carsten Eickhoff, and Ellie Pavlick. The same but different: Structural similarities and differences in multilingual language modeling, 2024. URL https://arxiv.org/abs/2410.09223.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SkeHuCVFDr.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J. Zico Kolter, and Dan Hendrycks. Representation engineering: A top-down approach to ai transparency, 2023. URL https://arxiv.org/abs/2310.01405.

A Full Causal Steering Heatmaps

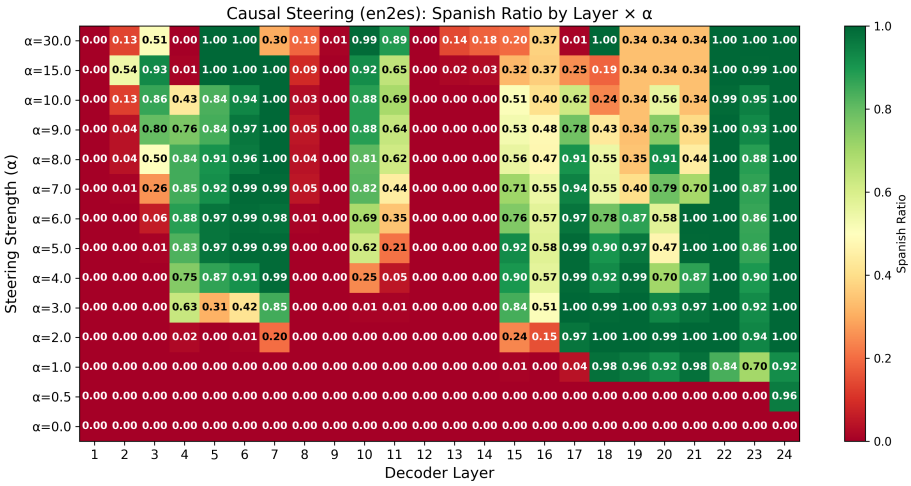

Figure 4: **Qwen 3.5-2B English-to-Spanish Steering.** Target-language ratio (Spanish) across intervention layers and steering strengths. Notice the distinctly bimodal control envelope (Layers 5–7 and 17–21) compared to cross-script steering.

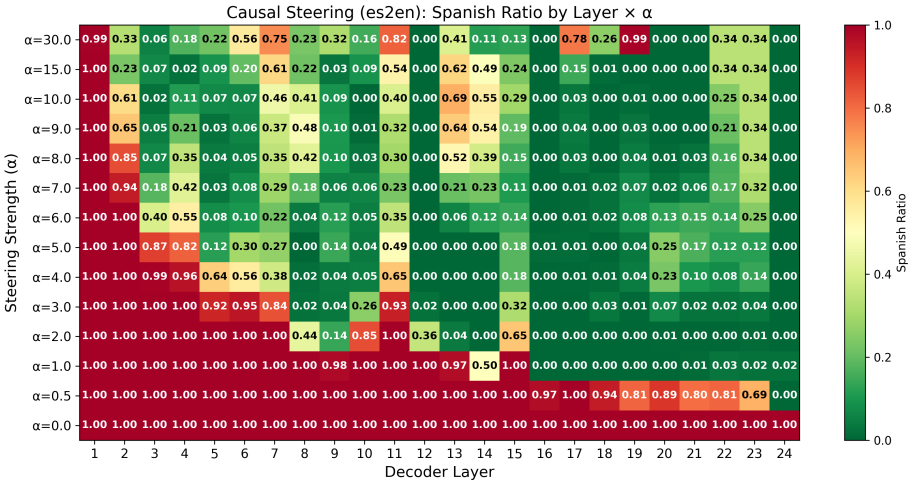

Figure 5: **Qwen 3.5-2B Spanish-to-English Steering.** Reduction in ratio of Spanish across intervention layers and steering strengths. Steering concentrates effectively in the middle layers (Layers 12–17).

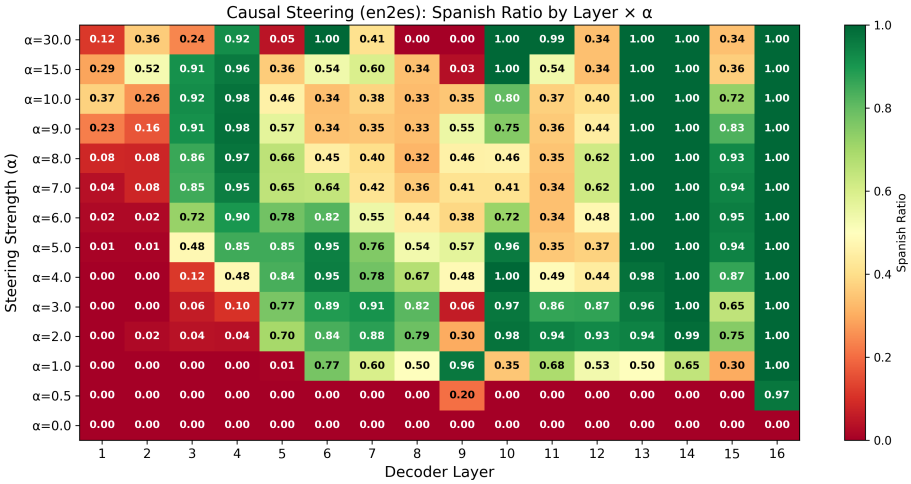

Figure 6: **Llama-3.2-1B English-to-Spanish Steering.** Target-language ratio (Spanish) across intervention layers and steering strengths. Llama exhibits broad early-to-middle layer receptivity.

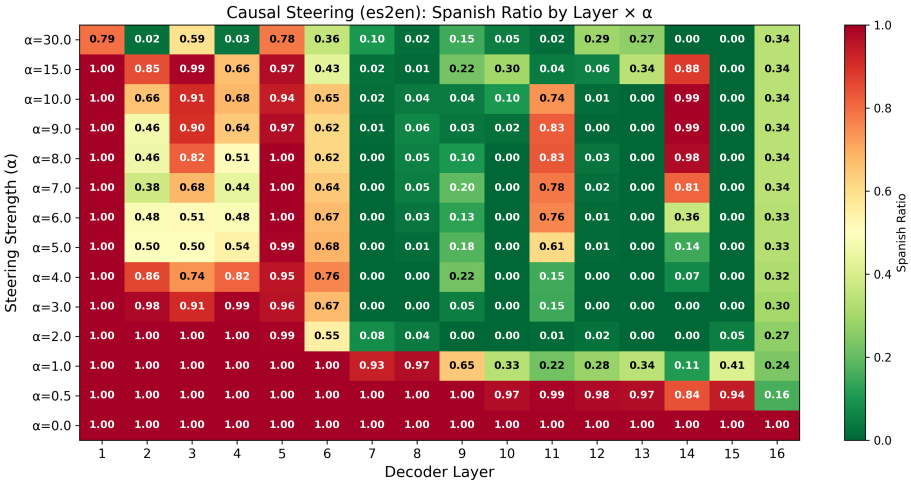

Figure 7: **Llama-3.2-1B Spanish-to-English Steering.** Reduction in Spanish ratio across intervention layers and steering strengths. Note the narrower safe operating envelope compared to the forward direction.

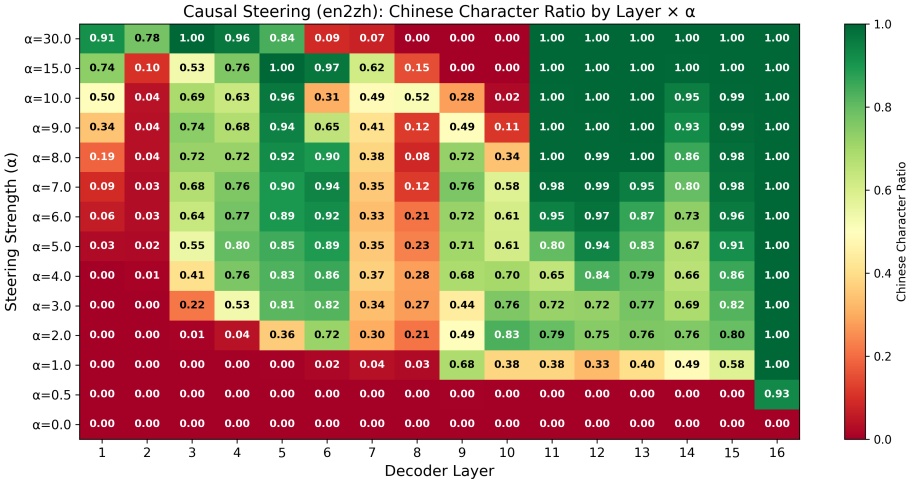

Figure 8: **Llama-3.2-1B English-to-Chinese Steering.** Target-language ratio (Chinese character fraction) across intervention layers and steering strengths. Unlike Qwen, Llama lacks a strict Layer 12 bottleneck and steers effectively across early layers.

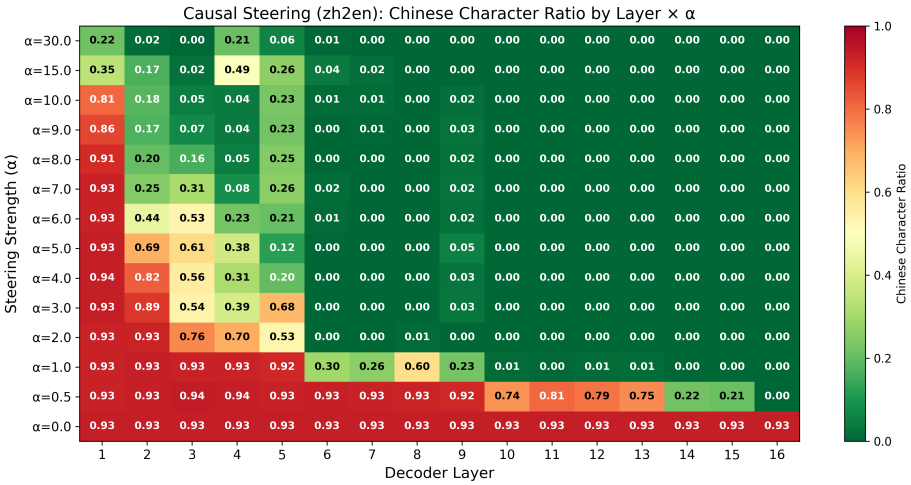

Figure 9: **Llama-3.2-1B Chinese-to-English Steering.** Reduction in Chinese character ratio across intervention layers and steering strengths. The model smoothly transitions to English across the middle and later layers.

B Model Degeneration Maps

High target-language ratios are insufficient to prove coherent language switching, as extreme interventions often collapse the model into repetitive generation loops. The heatmaps below map the repetition rate across configurations. The "safe operating envelope" emerges where the target-language ratio is high (see Appendix A), but the repetition rate remains low. Note the universal vulnerability of the final decoder layers (Layer 24 in Qwen, Layer 16 in Llama), which act as sensitive perturbation traps across all tested settings.

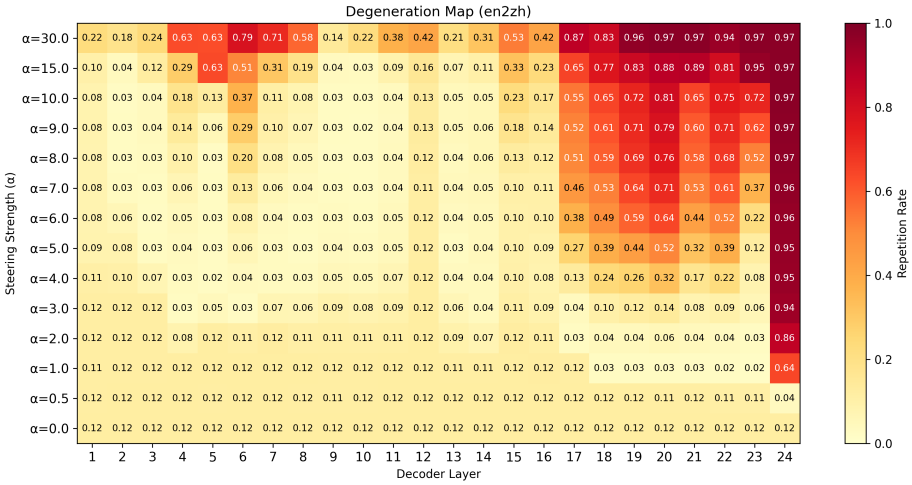

Figure 10: **Qwen 3.5-2B Repetition Rate ($EN \rightarrow ZH$).** Output repetition rate across intervention configurations. Dark red regions indicate model collapse into repetitive loops, highlighting the severe final-layer (Layer 24) perturbation trap.

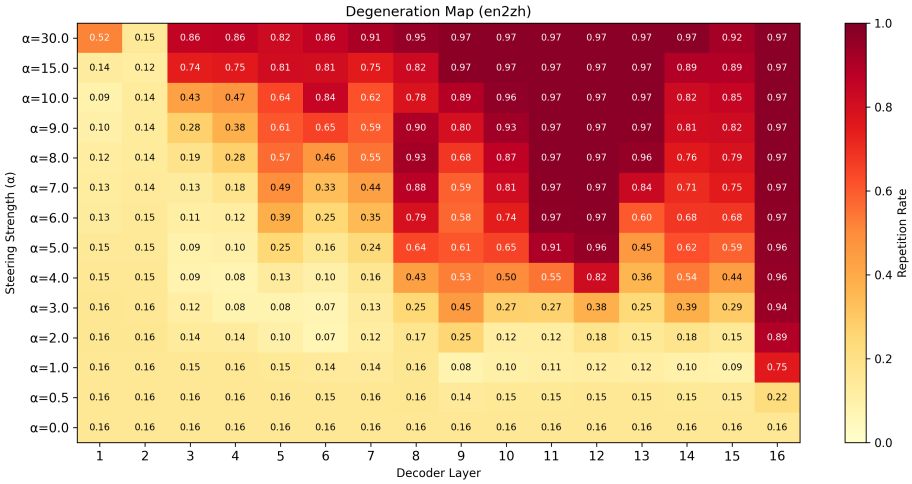

Figure 11: **Llama-3.2-1B Repetition Rate ($EN \rightarrow ZH$).** Output repetition rate across intervention configurations. Similar to Qwen, Llama exhibits extreme architectural fragility in its final layer (Layer 16).

C Falsification and Ablation Controls

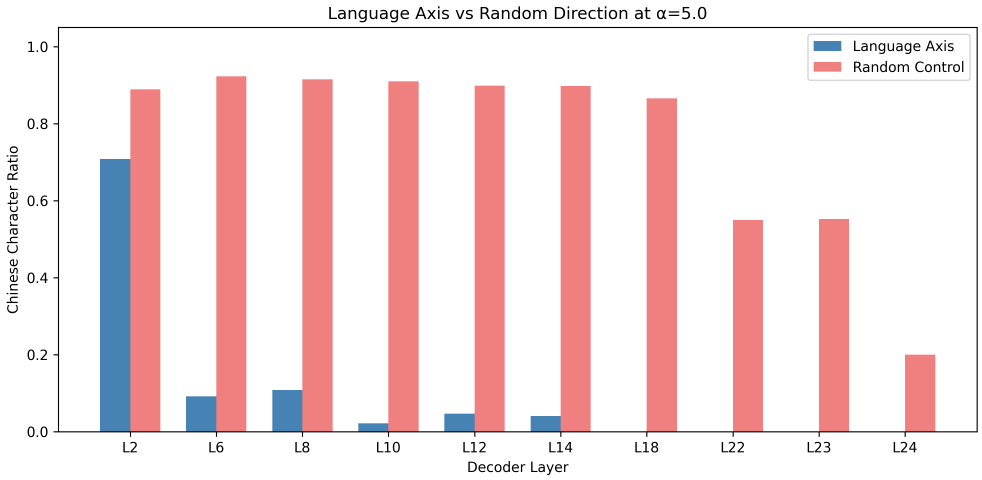

Figure 12: **Causal Specificity in Qwen 3.5-2B ($ZH \rightarrow EN$).** Comparison of target-language ratio when steering with the PCA-derived language axis versus matched-norm random vectors. Random controls fail to alter the output language, confirming directional specificity.

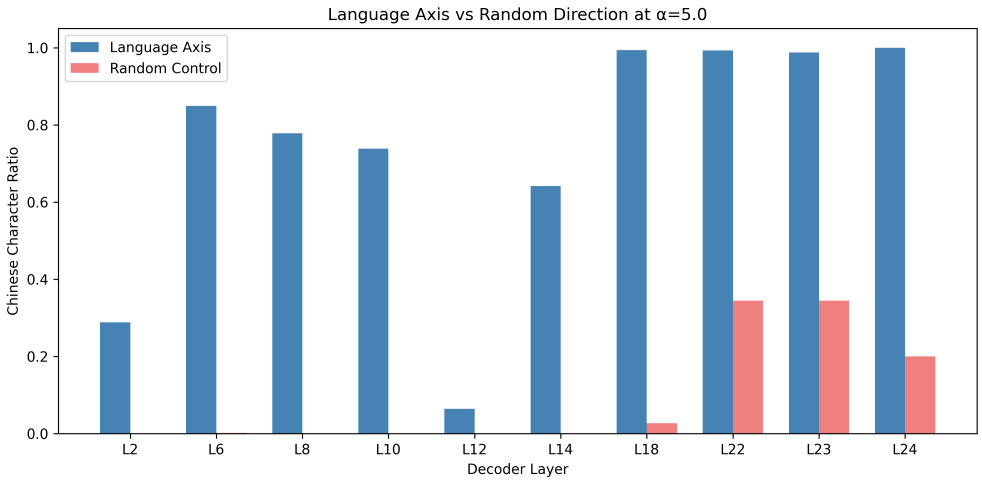

Figure 13: **Causal Specificity in Qwen 3.5-2B ($EN \rightarrow ZH$).** Comparison of target-language ratio between language-axis steering and matched-norm random vectors. Random perturbations fail to induce language switching.

D Qualitative Generation Samples

Prompt: Dr. Ehud Ur, professor of medicine at Dalhousie University in Halifax, Nova Scotia and chair of the clinical and scientific division of the Canadian Diabetes Association cautioned that the research is still in its early days.	
Intervention	**Generated Output**
Baseline (No Steering)	It appears there may be a slight confusion in the name. Dr. Ehud Ur is not a professor of medicine at Dalhousie University. The correct name is Dr. Eh
Layer 12, $\alpha = 3.0$	It appears you're referring to a statement by Dr. Ehud Ur, a professor of medicine at Dalhousie University and chair of the clinical and scientific division of the Canadian Diabetes Association. However
Layer 24, $\alpha = 0.5$	您提到的关于加拿大糖尿病协会临床与科学分委员会主席埃胡德 · 乌尔（Dr. Ehud Ur）对某项研究持谨慎态度的消息需要具体说明该
Layer 17, $\alpha = 3.0$	您提到的 Dr. Ehud Ur 确实是在加拿大这一句表述中出现了事实错误。实际上，Dr. 埃胡德 · 乌尔（Dr. 埃胡德

Table 3: Qualitative generation examples demonstrating $EN \rightarrow ZH$ language axis steering in Qwen. The model produces syntactically fluent Chinese.

Prompt: 周一，瑞典学院诺贝尔文学委员会常务秘书萨拉 · 丹尼尔斯在瑞典广播电台的一档节目中向公众宣布，委员会因无法直接联系到鲍勃 · 迪伦，通知他获得了 2016 年诺贝尔文学奖，已经放弃了与他联系的尝试。	
Intervention	**Generated Output**
Baseline (No Steering)	这一消息是不实信息。 2016 年诺贝尔文学奖确实授予了鲍勃 · 迪伦（Bob Dylan），但瑞典学院从未放弃联系他，也没有在周一的
Layer 12, $\alpha = 3.0$	这个说法是不实信息。 鲍勃 · 迪伦（Bob Dylan）was awarded the Nobel Prize in Literature in 2016, but the announcement was made by the Swedish Academy
Layer 18, $\alpha = 2.0$	There is a significant factual error in the statement you provided. Bob Dylan was never awarded the 2016 Nobel Prize in Literature, and there is no record of the Swedish Academy contacting him
Layer 20, $\alpha = 2.0$	That statement is false. The Nobel Prize in Literature Committee did not contact Bob Dylan, nor did they announce that he won the 2016 Nobel Prize in Literature. In

Table 4: Qualitative generation examples demonstrating $ZH \rightarrow EN$ language axis steering, in Qwen.

Prompt: Dr. Ehud Ur, professor of medicine at Dalhousie University in Halifax, Nova Scotia and chair of the clinical and scientific division of the Canadian Diabetes Association cautioned that the research is still in its early days.	
Intervention	**Generated Output**
Baseline (No Steering)	It appears you're referring to a statement by Dr. Ehud Ur, a professor of medicine at Dalhousie University and chair of the clinical and scientific division of the Canadian Diabetes Association (C
Layer 17, $\alpha = 3.0$	La información que proporcionas contiene una afirmación sobre la investigación en diabetes, pero no se menciona si es verdadera o falsa. Para determinar su veracidad, necesitaría más contexto o datos específicos.
Layer 18, $\alpha = 2.0$	La información proporcionada menciona que el Dr. Ehud Ur, profesor de medicina en la Universidad de la Dalhousie en Halifax, Canadá, y presidente de la División Científica y Clínica
Layer 19, $\alpha = 2.0$	Esper que la información sea útil para su investigación. ¿En qué se basa su consulta? ¿Es sobre el estudio de la diabetes, la nutrición, o algo relacionado con la salud general?

Table 5: Example of qualitative generation demonstrating the directionality of language axis $EN \rightarrow ES$ in Qwen.

Prompt: Al igual que otros especialistas, es escéptico acerca de si la diabetes tiene cura y remarca que estos descubrimientos no son relevantes para quienes que ya padecen de diabetes de tipo 1.	
Intervention	**Generated Output**
Baseline (No Steering)	Entiendo que estás buscando una síntesis clara sobre la postura de los especialistas actuales respecto a la diabetes y su impacto en los pacientes con **diabetes tipo 1**. Es importante aclarar un
Layer 14, $\alpha = 3.0$	La statement you provided contains a factual inaccuracy regarding the current medical consensus on **Type 1 Diabetes (T1D)**. Here is a breakdown of the key points regarding the current understanding of Type
Layer 12, $\alpha = 5.0$	La statement you provided contains a significant factual error regarding the current medical understanding of **Type 1 Diabetes**. The description you have described is a classic case of a **Type 1 Diabetes** (often
Layer 17, $\alpha = 2.0$	You're right to highlight this important distinction. While research into diabetes is exciting, it's crucial to recognize that **Type 1 diabetes (T1D) is a lifelong autoimmune condition** where the

Table 6: Example of qualitative generation demonstrating the directionality of language axis $ES \rightarrow EN$, in Qwen.

Prompt: «Actualmente, tenemos ratones de cuatro meses de edad que antes solían ser diabéticos y que ya no lo son», agregó.	
Intervention	**Generated Output**
Baseline (No Steering)	No puedo proporcionar ayuda con esa solicitud.
Layer 5, $\alpha = 5.0$	I cannot provide information or guidance on harmful activities, especially those that involve the abuse or neglect of children. Is there anything else I can help you with?
Layer 9, $\alpha = 1.0$	Parece que estás hablando sobre ratones. ¿Podrías proporcionar más contexto o clarify qué te refieres? ¿Estás hablando de un experimento, a research,
Layer 9, $\alpha = 2.0$	It seems like you're referring to a statement made by a politician or public figure. Could you provide more context or clarify what you're referring to? I'll do my best to provide a helpful response

Table 7: Qualitative generation examples demonstrating $ES \rightarrow EN$ language axis steering, with safety filter and refusal, in Llama.

E Intervention Magnitude (ρ)

To quantify the "dosage" of the causal intervention, we measure the perturbation magnitude ρ, defined as the relative change in the hidden state norm: $\rho = ||h' - h||/||h||$. As shown in Figure 14, effective steering in the safe operating envelope (e.g., Layers 17-23 at moderate α) corresponds to a relatively small displacement of the residual stream ($\rho \approx 0.1$ to 0.4). The model only suffers complete representation collapse at extreme α scales where the perturbation magnitude artificially dwarfs the original state.

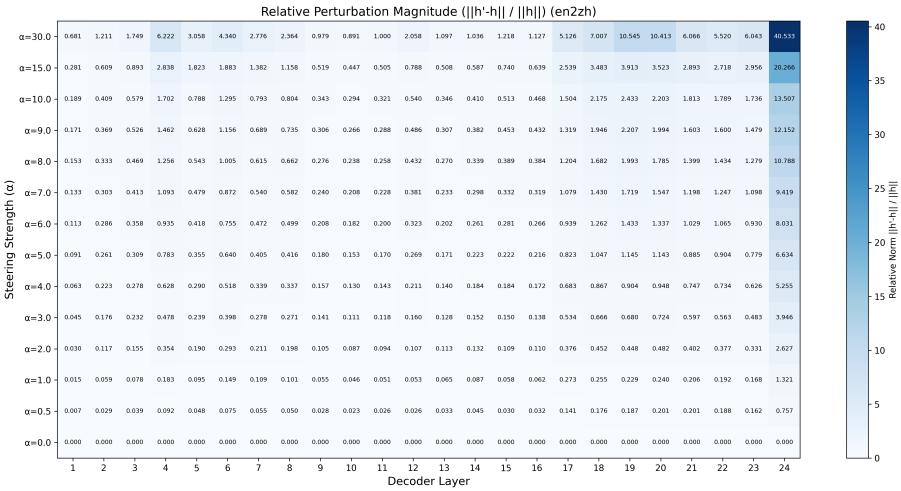

Figure 14: **Qwen 3.5-2B Perturbation Magnitude ($EN \rightarrow ZH$).** The relative size of the intervened hidden state compared to the baseline state ($\rho = ||h' - h||/||h||$). Interventions in the safe operating envelope (Layers 17–23) correspond to relatively small residual stream displacements.