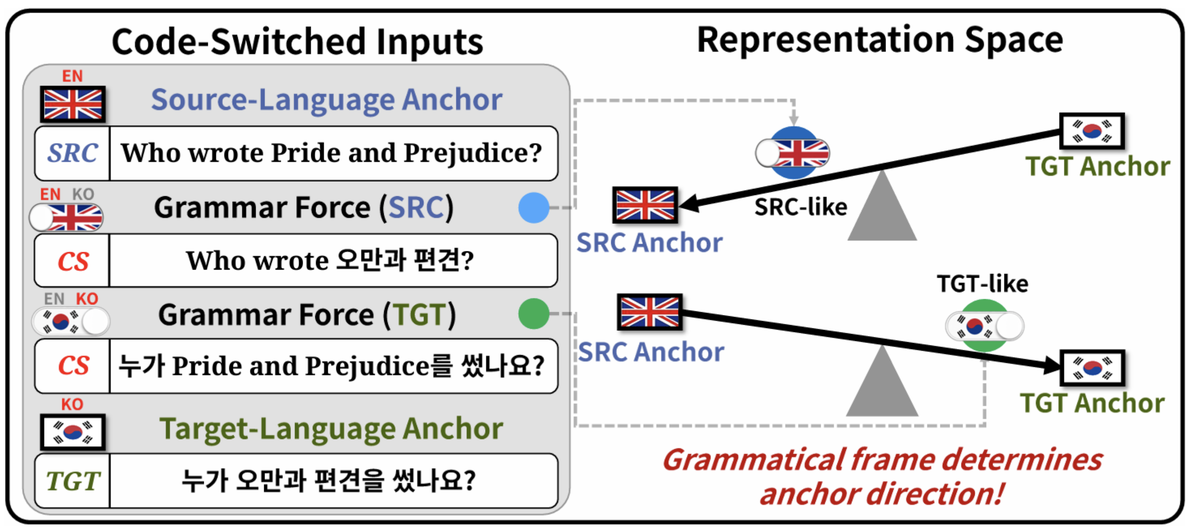
Code-Switching Reveals Anchor Bias in Multilingual LLMs

Jeonghyun Park, Seunghyun Yoon, Hwanhee Lee

Language Intelligence Lab, Chung-Ang University
Adobe Research, USA

ICML International Conference On Machine Learning

Motivation & Problem



Same information need, different grammatical frame

- Standard multilingual evaluation assumes **one language per input**.
- Code Switching separates **lexical mixing** from the sentence-level **grammatical frame**.
- If the frame dominates, the hidden state should move toward one language anchor.
- This matters because **the anchor predicts QA robustness**.

Anchor Bias

Deriving Raw Anchor Bias $H_M(q) \in \mathbb{R}^{L\times T\times d}$ **Hidden states** produced by model M for input q

$$\mathbf{r}_\ell(q) = \frac{1}{|C(q)|}\sum_{t\in C(q)} H_M(q)[\ell, t, :]$$ The layer-wise sentence **representation** obtained by averaging hidden states over **question** token span C(q)

$$s_\ell(q_i, q_j) = \cos\big(\mathbf{r}_\ell(q_i), \mathbf{r}_\ell(q_j)\big)$$ The cosine **similarity** between two inputs

$$\mathrm{AB}_\ell(q^{\mathrm{CS}}) = s_\ell\big(q^{\mathrm{SRC}}, q^{\mathrm{CS}}\big) - s_\ell\big(q^{\mathrm{TGT}}, q^{\mathrm{CS}}\big)$$

Raw Anchor Bias, computed as source-anchor similarity minus target-anchor similarity

Normalizing Raw Anchor Bias

$$\tilde{s}_\ell(q_i, q_j) = \frac{s_\ell(q_i, q_j) - \mu_\ell^M}{1 - \mu_\ell^M}$$ **Model- and layer-specific similarity** used to normalize raw cosine scores.

$$\widetilde{\mathrm{AB}}_\ell(q^{\mathrm{CS}}) = \tilde{s}_\ell\big(q^{\mathrm{SRC}}, q^{\mathrm{CS}}\big) - \tilde{s}_\ell\big(q^{\mathrm{TGT}}, q^{\mathrm{CS}}\big)$$ **Normalized** Anchor Bias after correcting for the layer-wise similarity (**anisotropy**).

Upper Half

$$\mathrm{AB}^{\mathrm{upper}}_{\mathrm{nrm}}(q^{\mathrm{CS}}) = \frac{1}{|\mathcal{L}_{\mathrm{upper}}|}\sum_{\ell\in\mathcal{L}_{\mathrm{upper}}} \widetilde{\mathrm{AB}}_\ell(q^{\mathrm{CS}})$$

Final Anchor Bias averaged over the upper layers

CS representations shift toward the frame language, predict QA degradation

Model	$\mathrm{AB}^{\mathrm{upper}}_{\mathrm{nrm}}$ GF-Src	GF-Tgt	Δ_{AB}	QA behavior (F1) SRC	GF-Src	GF-Tgt	SRC-anchored layers (%) GF-Src	GF-Tgt	Δ_{layer}
Dense models									
Aya-Expanse-8B	+0.300	-0.220	+0.520	9.3	7.0 (-2.3)	4.4 (-4.9)	85.6	42.8	+42.8
Qwen3.5-4B	+0.169	-0.262	+0.431	10.0	8.2 (-1.8)	7.2 (-2.8)	70.7	29.2	+41.5
Qwen3.5-9B	+0.185	-0.256	+0.441	13.1	11.9 (-1.2)	8.1 (-5.0)	72.9	32.5	+40.4
Qwen3.5-27B	+0.154	-0.197	+0.351	16.6	13.3 (-3.3)	10.4 (-6.2)	73.5	34.4	+39.1
Llama3.1-70B	+0.257	-0.225	+0.482	23.1	16.9 (-6.2)	13.5 (-9.6)	80.6	38.0	+42.6
Llama3.3-70B	+0.267	-0.191	+0.458	18.9	13.7 (-5.2)	11.6 (-7.3)	81.4	39.5	+41.9
MoE models									
Mixtral-8x7B	+0.273	-0.452	+0.725	10.8	6.7 (-4.1)	5.8 (-5.0)	74.6	33.2	+41.4
Phi-3.5-MoE	+0.248	-0.300	+0.548	13.1	9.6 (-3.5)	6.5 (-6.6)	64.8	31.7	+33.1
Qwen3-30B-A3B	+0.393	-0.081	+0.474	11.2	8.8 (-2.4)	6.5 (-4.7)	78.5	42.1	+36.4
Qwen3.5-35B-A3B	+0.225	-0.214	+0.439	21.5	15.9 (-5.6)	10.4 (-11.1)	74.3	37.5	+36.8
Mean	+0.247	-0.240	+0.487	14.8	11.2 (-3.6)	8.5 (-6.3)	75.7	36.1	+39.6

Model	Type	Random cos.	$1-\mu$
Aya-Expanse-8B	Dense	0.7944	0.2056
Llama3.3-70B	Dense	0.8729	0.1271
Llama3.1-70B	Dense	0.8775	0.1225
Qwen3.5-4B	Dense	0.9112	0.0888
Qwen3.5-27B	Dense	0.9145	0.0855
Qwen3.5-9B	Dense	0.9320	0.0680
Mixtral-8x7B	MoE	0.8862	0.1138
Phi-3.5-MoE	MoE	0.8698	0.1302
Qwen3.6-35B-A3B	MoE	0.9379	0.0621
Qwen3-30B-A3B	MoE	0.9791	0.0209

Random CosSim caused by anisotropy

Raw cosine similarity can overstate representational closeness!

- Internal **anchor bias follows the input's grammatical backbone** rather than its lexical mix: **GF-SRC remains source-oriented** with a positive mean bias of +0.247, whereas **GF-TGT shifts target-ward** with a negative mean bias of −0.247.
- This **frame-dependent anchor shift aligns with QA behavior**: GF-SRC, which stays closer to the source anchor, incurs a smaller average F1 drop than GF-TGT (−3.5 vs. −6.2), suggesting better preservation of task performance.
- This **anchoring pattern persists across depth**, with GF-SRC source-anchored in 75.7% of layers versus only 36.1% for GF-TGT, showing that grammatical frame robustly shapes internal representations.