Whose Taiwan? Cross-Lingual Geopolitical Framing in the Representations of Large Language Models

Guan-Yu Tseng Lang-Ching Yeh Yu-Chieh Wang Shu-Kai Hsieh
Graduate Institute of Linguistics, National Taiwan University
{r14142007, r14142008, r14142006, shukaihsieh}@ntu.edu.tw

Abstract

Evidence that large language models hold skewed views of contested geopolitical entities usually comes from what they generate, which shifts with the prompt and is easy to over-read. We ask the question inside the model instead: does it bring a different frame to "Taiwan" when the surrounding sentence is English rather than Taiwan Mandarin? We pair sentences that say the same thing in the two languages, add Japan and Iceland as controls, and use Natural Language Autoencoders to turn the model's internal representation of "Taiwan" into readable text. In both Qwen2.5-7B and Gemma-3-12B, an English sentence makes that representation read "Taiwan" through a China-centered geopolitical lens nearly five times as often as a Mandarin one (14.3% vs. 3.0%), with no such gap for the controls. What changes is *which frame* the representation falls into, not *how far* it moves: the language switch displaces "Taiwan" no more than it displaces the controls. The frame can also be switched on directly, clearly in Gemma and directionally in Qwen: adding a direction computed from our own stimulus labels, which the readout never sees, pushes a Mandarin-context representation into the China-anchored state, more strongly at higher doses and back below baseline when subtracted, while equally large random directions do nothing. This holds in the representation our method reads, not yet in the text the model writes. Still, the language a user writes in can quietly change which frame a model brings to their own cultural referents.

Keywords: Mechanistic Interpretability, Cultural Bias, Sovereign AI, Frame Semantics, LLMs

1 Introduction

Large language models are pre-trained on corpora dominated by a few high-resource languages, predominantly English and Chinese, and their behaviors and internal representations inherit the cultural and ideological paradigms of those majority datasets (Wendler et al., 2024; Schut et al., 2025). Behavioral probing via targeted prompting is increasingly used to test for such cultural biases (Tao et al., 2024; Kovač et al., 2023), and the worldviews a model encodes are a critical dimension of AI safety, since skewed outputs can subtly shape user cognition (Agarwal et al., 2025). The issue is acute for Taiwan: most Chinese-capable LLMs are trained overwhelmingly on Simplified-Chinese corpora from mainland China, inheriting that region's cultural and geopolitical defaults (Lyu et al., 2025) and marginalizing Taiwan's identity and Traditional-Chinese usage, the data hegemony that motivates sovereign AI.

In this work we move from behavior to representation. Using NLA, we ask whether, holding propositional content fixed, the residual-stream representation of "Taiwan" falls into different semantic frames depending on whether the surrounding context is English or Taiwan Mandarin. A code-switched condition aims to separate context language from orthographic form; matched control entities (Japan, Iceland) separate Taiwan-specific effects from generic properties of English text; and two models (Qwen2.5-7B, Gemma-3-12B), each read out by its own verbalizer, test model-dependence while guarding against self-interpretation circularity. To our knowledge this is the first study of representational bias toward Taiwan at the mechanistic rather than behavioral level.

We find that English contexts anchor "Taiwan" to a China-centered geopolitical frame about five times more often than Mandarin contexts, concentrated at the mention token, specific to Taiwan, and carried by English wording rather than simplified orthography (in both models, more pronounced in Qwen). Our central claim is that this is a differ-

Proceedings of the 38th Conference on Computational Linguistics and Speech Processing (ROCLING 2026), pages 136–150
November 6-7, 2026 ©2026 Association for Computational Linguistics

ence in *which frame* the representation occupies, not in *how far* it moves under the switch: the same displacement carries no Taiwan-specific geometric signature, so a distance-based probe is silent where a frame-level readout is not. A steering analysis then shows the frame to be a manipulable direction, and locates the carrier in a metadata-defined frame direction rather than in the language difference itself, established in the representation the verbalizer reads rather than the model's generated text.

2 Literature Review

2.1 Behavioral Probing of Cultural Bias and Its Ceiling

A large body of work evaluates the cultural and ideological dispositions of LLMs through their *outputs*, comparing model responses on value surveys and culturally grounded prompts against human populations (Tao et al., 2024; Durmus et al., 2023; Ramezani and Xu, 2023; Arora et al., 2023; Cao et al., 2023), establishing that models carry measurable, often Western- and English-leaning defaults that shift with the language of the prompt, a single model becoming a "superposition" of cultural perspectives (Kovač et al., 2023).

Yet the behavioral paradigm faces two limits: prompt-based probes measure surface behavior under a specific elicitation, sensitive to wording and maskable by alignment, so they cannot localize whether a disposition resides in the representation or the response policy; and cultural readings of model text may invite the charge of *over-reading*. Both motivate a move to representation with a method that attaches an explicit reliability score to every interpretation.

2.2 Mechanistic Interpretability and Verbalizing Internal Representations

Mechanistic interpretability instead explains behavior in terms of the internal computation over the residual stream. One family decomposes activations into human-interpretable features via sparse autoencoders (Cunningham et al., 2023; Bricken et al., 2023; Templeton et al., 2024); a second projects hidden states into the vocabulary, as in the logit and tuned lens (Belrose et al., 2023); and a third, most relevant here, uses the model *itself* to translate a hidden state into natural language, as in Patchscopes (Ghandeharioun et al., 2024) and SelfIE (Chen et al., 2024).

These make representations legible but give no intrinsic measure of how *faithful* a verbalization is to the vector it describes, a property directly relevant to the over-reading objection. Natural Language Autoencoders (NLA) (Fraser-Taliente et al., 2026) supply it: an *activation verbalizer* (AV) maps a residual-stream vector to free text and an *activation reconstructor* (AR) maps it back, so the round-trip error is a *faithfulness score*: a low error means the description alone recovers the original direction rather than an analyst's projection.

This back-translation check, absent in other methods, is a necessary though not sufficient condition for trusting a reading, and is well-suited to a contested domain such as cultural semantics; recent work nonetheless cautions that verbalization methods may not convey information privileged to the activation itself (Li et al., 2025), which is why we treat the score as auditable evidence rather than proof. NLA also complements representation steering, which shows that activation-space directions can causally control generation (Turner et al., 2024; Rimsky et al., 2024; Zou et al., 2023).

2.3 Cross-lingual Representation, the English Pivot, and the Sovereign-AI Gap

A parallel line of interpretability work asks what language a multilingual model "thinks in." Wendler et al. (2024) show that English-heavy transformers route non-English inputs through an English-centric latent space in intermediate layers, and Schut et al. (2025) report that multilingual models compute over English-aligned representations even when prompted and answering in other languages (see also Zhao et al., 2024).

If such a pivot is tilted toward one language's defaults, the same proposition in two languages need not land in the same internal frame. That matters most for entities with contested cultural status, the concern behind Taiwan's sovereign-AI models and Traditional-Chinese evaluation suites (Lin and Chen, 2023; Tam et al., 2024). "Taiwan" is such an entity: whether a model activates an in-group everyday-life frame or a China-anchored (geopolitical) frame is a direct measure of whose worldview its representation encodes. The closest prior work scores model *answers*: Ko (2026) finds that one model states different positions on Taiwan's sovereignty in Chinese than in English, and Hsieh et al. (2024) benchmarks social bias through a Taiwan cultural lens. Such biases are thus documented behaviorally; to our knowledge none has measured

them at the level of the entity's residual-stream representation, which is the gap this paper addresses.

3 Data and Methodology

We study *within-model* representations: for the same proposition realized in English and in Taiwan Mandarin, we extract the residual-stream activation at the "Taiwan" ("台灣") token, verbalize it with a natural-language autoencoder, score each verbalization by its round-trip faithfulness, and annotate every description for the semantic frame, geopolitical anchoring, and construal it exhibits.

3.1 Research Questions and Predictions

Treating the AV description as a readable projection of the representation, we organize the central question into two research questions.

RQ1: Existence and specificity. We predict that **(P1a)** an English context frames "Taiwan" more geopolitically, foregrounding sovereignty even for neutral, non-political stimuli, whereas a Mandarin context leans toward an everyday, in-group reading (daily life, local culture, geography); **(P1b)** this English-over-Mandarin gap is *specific to Taiwan*, weak or absent for the control entities Japan and Iceland; and **(P1c)** an English context construes "Taiwan" more as a political actor or polity, a Mandarin context more as a place, people, or lived experience.

RQ2: Nature, locus, and cause. We predict that **(P2a)** the shift is an anchoring *overlay* rather than a change of topic: the exhibited topical-frame distribution is largely language-invariant while the geopolitical anchoring laid over it differs; **(P2b)** it is driven primarily by *context language* rather than the mention's script (台灣 vs. "Taiwan"), which a code-switching design can dissociate; and **(P2c)** if anchoring is a causal property of the representation, a direction separating geopolitically framed from non-geopolitical items should, when added to an unanchored (Mandarin-context) vector, induce the anchored reading in a dose-dependent and sign-reversible way that norm-matched random directions do not.

3.2 Minimal-Pair Corpus

Factorial design. The corpus crosses a manipulated *context language* (English vs. Taiwan Mandarin) with a stratifying *stimulus frame*, an *entity* factor for difference-in-differences control, and, in a dedicated sub-experiment, the *orthographic form* of the mention (Appendix Table 2). Frames follow FrameNet as domain-level scenes (Fillmore, 1982; Baker et al., 1998); each sentence evokes exactly one frame, does *not* mention China, and (for core items) stays evaluatively neutral. Crucially, the international-status frame (POL-INT) is deliberately de-Sinicized: it discusses international standing without naming China, so that any China reference in a verbalization is a representation-level *insertion* rather than a surface echo of the stimulus.

Template grammar. Every sentence enforces a three-part structure: *lead-in* + *target clause (entity as subject)* + *continuation*. The lead-in establishes the frame in ≥ 10 tokens without naming the entity, since in an autoregressive model a sentence-initial entity would deprive Site A of causal context from the language manipulation, and the continuation ensures Site B carries the full propositional gist.

The two language versions are back-translation-verified translation equivalents, rated for naturalness on a 1 to 5 scale (items below 4 rewritten) with token-length difference $\leq 20\%$. The entity may not appear as a proper-name modifier (for example "Taiwan Strait"), whose vector would not represent the entity itself.

Controls and Design B. To ensure an English-Mandarin difference is Taiwan-specific rather than a general property of English text, we add two control entities via minimally contrastive slot templates, *Japan* (high visibility, no sovereignty dispute) and *Iceland* (a geopolitically bland baseline), with the language $\times$ entity interaction as the test statistic. *Design B* is a 2×2 code-switching sub-experiment crossing a Latin "Taiwan" and a Hanzi "台灣" against both context languages; code-switched inputs are mildly out-of-distribution for the fineweb-trained NLA, reconstructing 1.3 to $1.4\times$ worse than baseline, so Design-B conclusions are read with that caveat (§3.4).

Scale. The corpus comprises 360 sentences: 192 core Taiwan items, 120 control items, and 48 Design-B items, with the full factorial breakdown in Appendix Table 3. Two sites $\times$ k=5 verbalizations yields 3,600 descriptions per model, 7,200 across the two.

3.3 Target Models and Activation Extraction

We study two instruction-tuned open-weight models with official NLA checkpoints, run in parallel to

guard against the circularity of interpreting a model with a verbalizer derived from itself: **Qwen2.5-7B-Instruct** (Qwen Team, 2024) (layer 20 of 28) and **Gemma-3-12B-IT** (Gemma Team, 2025) (layer 32 of 48), with the extraction layer for each model read from checkpoint metadata.

Following the NLA setup we feed raw text without any chat template, keeping activations in-distribution, and record two vectors per sentence: *Site A*, the last subtoken of the mention (where autoregressive information about the entity aggregates), and *Site B*, the final non-special token (a sentence-level gist).

3.4 Natural-Language Autoencoding and the Faithfulness Gate

We verbalize each vector with the NLA activation verbalizer (Fraser-Taliente et al., 2026), drawing $k = 5$ samples at temperature 0.8 to estimate the stability of the description distribution. Each description is scored by the activation reconstructor, which reads only the text and predicts a vector; round-trip faithfulness is

$$\mathrm{MSE} = 2\,(1 - \cos(\hat{v}, v)) \in [0, 4], \qquad (1)$$

lower being more faithful.

We use this score to make *reliability* auditable rather than to filter the corpus: the primary analysis retains all descriptions, and the score plays two roles. First, it validates the verbalizer, since a low round-trip error confirms a description recovers its vector rather than an analyst's projection. Second, the *faithfulness gate* marks the most-faithful within-model MSE tertile, within which we verify that the framing signal survives (§4).

3.5 Framing Annotation Scheme and Reliability

The scheme describes *how* an entity is conceptualized rather than merely its polarity, so it codes categories rather than sentiment. Its vocabulary is borrowed from frame semantics, in which understanding a word invokes a structured background scene (Fillmore, 1982; Baker et al., 1998), and from cognitive grammar's notion of construal, the same referent profiled as lived experience, as an object of analysis, or as a political actor (Langacker, 1987).

The unit of annotation is a single AV description, and the object is which frame *the text itself* exhibits, not whether it correctly reads the stimulus (a festival verbalized as "dessert" is coded FOOD). Annotators see only the AV text when coding frames; the stimulus is revealed only for the drift covariate. The scheme has five dimensions (Appendix Table 5): exhibited frame (D1), anchoring (D2), construal (D3), orthography (D4), and drift (D5).

D2 is coded with four sub-labels and analyzed as binary, ANCHORED $= \{$PRC, GEOPOL-OTHER$\}$ versus unanchored (the pilot found the IN-GROUP/NEUTRAL boundary unreliable and irrelevant to the anchoring analysis, so both fall under "unanchored"), with PRC (explicit China-anchoring) also reported alone. "Geopolitical intrusion" is a non-political stimulus frame co-occurring with D1=POL-INT or an anchored D2.

Reliability and validity. The codebook was frozen through a double-blind human pilot (two annotators, 80 items), which required prevalence-adjusted agreement ≥ 0.7 on the gating dimensions D1 and binary D2 (D1 raw $\kappa = 0.83$, clearing it without adjustment; D2 PABAK $= 0.85$, prevalence-adjusted because anchoring is rare and depresses raw κ; App. C).

The full pool of 7,200 descriptions was then triple-annotated by three independent LLM annotators (Gilardi et al., 2023; Ziems et al., 2024). We judge reliability against the human ceiling rather than an absolute threshold: inter-annotator agreement meets or exceeds the pilot on every dimension (Appendix Table 6), following standard inter-coder practice (Cohen, 1960; Landis and Koch, 1977; Artstein and Poesio, 2008). Reliability (agreement) is distinct from validity (match to human judgment), which we assess on a separate stratified 240-item sample, disjoint from the reliability set, re-annotated by two authors under the pilot protocol (§4).

3.6 Causal Validation: Frame-Direction Steering

To test P2c we intervene directly: for a test vector h and candidate direction δ we verbalize $h + \alpha\delta$ and annotate the result for D2 under the identical codebook and pipeline, so the causal and descriptive readouts measure the same construct.

Directions. All directions are difference-in-means estimates, the standard steering-vector construction (Zou et al., 2023; Rimsky et al., 2024). The primary direction Δ_{frame} contrasts the mean activation of geopolitically framed stimuli (POL-INT, HIST) with that of non-geopolitical stimuli

within a single context language; defined purely from stimulus metadata the verbalizer never observes, it makes steering plus AV+D2 a genuine cross-construct test.

Three comparisons complete the design: Δ_{anchor} (high-minus-low over the D2 labels, sharing its construct with the readout, so reported only as an *instrument check*); Δ_{lang}, the English-minus-Mandarin displacement of §4.2, as a negative reference; and five norm-matched random directions as a null. Every Δ is re-estimated under leave-one-frame-out (LOFO), so no vector is pushed along a direction its own frame helped define.

Dose and estimand. Test vectors are the Taiwan Site A, Mandarin-context cells, the *low*-anchoring pole, so the intervention asks whether the representation can be moved into the anchored state an English context produces on its own. Since α is not comparable across models we sweep the relative displacement $\|\alpha\delta\|/\|h\|$ (Li et al., 2023), and the estimand is the per-vector paired difference in ANCHORED rate against the unsteered (α=0) condition (Vig et al., 2020), with 95% bootstrap CIs.

Of the four pre-specified reading criteria two carry most of the weight: a null for the norm-matched random directions, which excludes a pure magnitude artifact, and sign reversal under $\alpha < 0$, the standard for attributing a behavior to a single direction (Arditi et al., 2024). Test-set sizes, dose grids, and frame groups are in Appendix E.

3.7 Analytical Approach

Our confirmatory tests are pre-specified. RQ1 was designed around the mixed-effects logistic regression `anchored ~ language * entity + (1|template) + (1|frame)`, whose crossed random effects generalize beyond the specific stimuli (Baayen et al., 2008; Barr et al., 2013). Control anchoring sits at the floor, however, so the language × entity interaction separates, and we report P1b as a descriptive difference-in-differences (DiD) with 95% item-clustered bootstrap CIs, P1a as a within-Taiwan cluster-robust logistic regression, and P1c as a χ^2 test of independence (Cramér's V).

Two annotation-independent checks complement the labels: a stimulus-frame × exhibited-frame drift matrix compared by Jensen-Shannon divergence, and a geometric profile of Δ_{lang} computed after per-dimension standardization (Sun et al., 2024). Three threats to validity (output-language confounding, self-interpretation circularity, and the out-of-distribution status of code-switched items) are addressed by design, with full specifications in Appendix E; a fourth, cross-language difference in the mention's surprisal, is not controlled and is noted in §6.

4 Results

Before the analysis we verify the pipeline (full tables in Appendix C). Round-trip faithfulness is high (median MSE 0.009 for Gemma, 0.193 for Qwen), so the verbalizer reconstructs the vectors it describes; the primary analysis uses all 7,200 descriptions, and the language effect below also holds inside the most-faithful within-model MSE tertile (+6.5 points, 95% CI $[+3.4, +9.8]$, item-clustered bootstrap).

The annotation is reliable and valid: across three independent LLM annotators inter-annotator agreement meets or exceeds the human pilot on every dimension (frame Fleiss $\kappa = 0.926$; binary anchoring 0.773; Table 6), and on a disjoint, human-validated 240-item sample model-human agreement is comparable to the human-human benchmark for D1-D4 (Table 7); only drift (D5) falls short of an already-low ceiling and serves as a descriptive covariate only.

4.1 English Contexts Anchor "Taiwan" to China (RQ1)

Context language has a large effect on geopolitical anchoring (Table 1, Appendix Fig. 1). Across the 4,320 Taiwan baseline descriptions an English context produces an explicit China-anchored reading (D2=PRC) in 14.3% of verbalizations versus 3.0% in a Mandarin context, a nearly fivefold difference. The same asymmetry holds for any geopolitical anchoring (within-Taiwan odds ratio 3.4, cluster-robust p=5×10^{-10}) and for geopolitical intrusion into non-political stimulus frames, and it is unchanged when the verbalizer's own output language is held fixed (14.3% versus 3.4% within English-language descriptions), so it is not an artifact of which language the description happens to be written in.

The effect is sharply localized: it is roughly three times larger at the mention token than at the sentence-final token, so the anchoring sits where information about the entity aggregates rather than diffusing across the sentence.

Measure	En	Zh	Test
China-anchoring (PRC)	14.3	3.0	V=.20, p<10^{-39}
at Site A (mention)	21.2	3.8	
at Site B (sentence)	7.4	2.1	
Any anchoring	17.8	6.0	V=.18, p<10^{-32}
Geopolitical intrusion	10.1	1.5	V=.19, p<10^{-22}
Construal = ACTOR	8.8	6.7	V=.05, p=.006
Construal = EXPER.	26.4	29.5	

Table 1: Cross-lingual contrasts for Taiwan baseline descriptions (% of descriptions), English vs. Mandarin context. Intrusion is computed over non-political stimulus frames.

4.2 The Effect Is Taiwan-Specific in Annotation, but Not in Geometry (RQ1)

The control entities isolate what is specific to Taiwan (Appendix Fig. 2): Japan and Iceland are almost never anchored in either language ($\leq 0.2\%$), whereas Taiwan swings from 6.0% to 17.8%. Because the controls sit at the floor in both languages, the difference-in-differences in anchoring rate is carried almost entirely by Taiwan: $+11.8$ points (95% CI $[+8.8, +15.0]$, item-clustered bootstrap), so geopolitical anchoring is essentially Taiwan-only. For the same reason the formal language $\times$ entity interaction is estimated imprecisely (near-complete separation), and we rest P1b on this descriptive DiD and the within-Taiwan contrasts.

Geometry yields a different result. Measuring $\|\Delta_{\text{lang}}\|$ in standardized space (Appendix Fig. 6, App. F), Taiwan does *not* move more under a language switch than the controls: for Qwen displacements are near-identical across entities (DiD in displacement magnitude $+1.7$, Site A, n.s.), and for Gemma the only significant contrast runs the *other* way (Taiwan 58.3 vs. control 67.5, DiD -9.2 in the same units).

Annotation supports P1b while geometry detects no Taiwan-specific language effect: evidence that the effect is a property of *which frame* the same-sized displacement lands in, not of *how far* the vector moves.

4.3 Frame Drift and Construal (RQ2)

The stimulus-to-exhibited-frame drift matrix (Appendix Fig. 3) is strongly diagonal, confirming a stable pipeline, and between languages the frame-identity distributions barely differ (mean row Jensen-Shannon divergence 0.015), locating the cross-lingual effect not in *which topic* the description is about but in the *geopolitical framing* laid over it.

Consistently, anchoring concentrates in two frames, POL-INT (46.7%) and HIST (20.8%, almost all PRC), while the other six stay below 8% and act as clean baselines. Construal (P1c) shifts as predicted but weakly: English descriptions construe Taiwan slightly more as a political ACTOR (8.8% vs. 6.7%) and less experientially (26.4% vs. 29.5%; χ^2=10.1, p=.006, V=0.05).

4.4 Qwen Is More Sinocentric and Less Stable than Gemma

The two models differ sharply and consistently (Appendix Fig. 4). On Taiwan baseline descriptions Qwen produces far more China-anchoring than Gemma (PRC 15.0% vs. 2.3%; any anchoring 18.1% vs. 5.7%; model$\times$anchoring χ^2=155, V=0.19), more geopolitical intrusion (11.2% vs. 0.4%), more frame drift (47.6% vs. 18.7%), and more simplified orthography (23.2% vs. 2.0%; χ^2=511, V=0.49): the more heavily mainland-trained model is both more likely to impose a China-centered frame and less stable, and since each model is read out by its own verbalizer this cannot be a shared-interpreter artifact.

One caveat bounds the gap's *magnitude*, not its direction: Qwen's median round-trip MSE is $\sim$20$\times$ Gemma's (0.193 vs. 0.009), so part of Qwen's higher instability, frame drift especially, may reflect noisier readout rather than a more Sinocentric representation. We therefore rest the comparison on the consistent *direction* across every China-leaning index and treat the cross-model *size* as directional, not calibrated.

4.5 Orthography and Sentence-Level Locus (RQ2)

China-anchoring does *not* travel with simplified-Chinese orthography: of the China-anchored descriptions the overwhelming majority are English (350 vs. 23 mixed), and within Taiwan items more PRC anchoring occurs in Traditional-orthography descriptions (n=48) than simplified ones (n=16). It is carried by English geopolitical vocabulary (*Chinese Taipei*, *province*, *cross-strait*), while simplified orthography is a separate, largely Qwen-specific tendency, so D4 and D2 are decoupled and simplified characters should not proxy for a China-anchored representation.

The code-switching sub-experiment locates the effect: at Site B the language displacement far exceeds the script displacement, so sentence-level representation is driven by context language (P2b); at

Site A the two are comparable (for Gemma, script is larger), consistent with the local orthographic signal decaying as context language accumulates downstream.

4.6 Steering a Frame Direction Causally Induces Anchoring (RQ2)

Adding Δ_{frame} to unsteered Mandarin-context vectors reproduces the anchoring effect, and all four pre-specified criteria are met in Gemma (Appendix Table 10, Appendix Fig. 5).

Dose-response and reversal. In the two political frames the anchoring rate rises monotonically with dose: from 0% at the strongest negative dose through 22.2% unsteered to 75.0% at the strongest positive dose (per-vector Spearman $\rho{=}0.58$, $p{=}4{\times}10^{-11}$; paired difference at the top dose $+0.53$, 95% CI $[+0.25, +0.78]$), and reverses: negative doses drive anchoring *below* baseline (-0.22, $[-0.39, -0.08]$), so this is not a generic "perturbation degrades the description" artifact. The label distribution shifts gradually rather than collapsing (INGROUP $70\% \rightarrow 39\%$, GEOPOL-OTHER $3.5\% \rightarrow 37\%$), with the explicit PRC sub-label rising in parallel ($2.1\% \rightarrow 20.8\%$; paired $+0.19$, $[+0.10, +0.28]$).

Random null and generalization. Norm-matched random directions produce no comparable movement: over five seeds their effect stays within ± 0.11 at every dose with no dose-response, against $+0.53$ for Δ_{frame} ($+0.54$, $[+0.44, +0.63]$ when each steered vector is contrasted against its own random average).

Crucially the effect generalizes beyond the frames that defined the direction: GEO vectors, excluded under LOFO and never anchored at baseline (0%), reach 66.7% ($\rho{=}0.70$, $p{=}3{\times}10^{-9}$), and the *cold* frames rise from 0% to 48.9%; since these are the cleanest, lowest-baseline cells, their movement is the least ambiguous evidence that the readout responds linearly to this direction rather than merely amplifying already-geopolitical vectors.

The language direction is not the causal carrier. Δ_{lang}, the direction our descriptive geometry recovered, does *not* behave this way: in Gemma it leaves the cold frames untouched (0.00 throughout) and GEO nearly so (≤ 0.06), moving only the already-geopolitical frames (peak $+0.33$), and in Qwen it is non-significant everywhere. This is the same null §4.2 found geometrically, now confirmed causally: the English-minus-Mandarin displacement is a broad pan-language direction, not the geopolitical frame. P2c is thus supported in general (an anchoring direction exists and is manipulable), but the causal work is done by the frame-metadata direction, not the language direction.

Qwen is directionally consistent but underdosed. Qwen replicates the effect in direction but weaker, only at the top of its grid: hot-frame anchoring rises from 34.7% to 50.0% with a significant monotone trend ($\rho{=}0.30$, $p{=}2{\times}10^{-4}$) and a significant held-out GEO effect ($+0.14$, $[+0.03, +0.28]$), but no cold-frame movement and no clean reversal (per-vector CIs in Appendix Table 10). Gemma's hot-frame effect is roughly three times Qwen's at a matched dose, a contrast we report cautiously given Qwen's higher baseline and unverified dose linearity.

5 Discussion

The frame, not the distance. Context language changes *which frame* the model brings to "Taiwan" while leaving the raw geometry of the language switch largely unchanged. The annotation line is unambiguous: an English context makes a China-anchored reading about five times more likely, concentrated at the mention token and absent for controls. Yet the geometric line finds no Taiwan-specific enlargement of the displacement, and for Gemma it is even smaller, so the same-sized shift lands in a more Sinocentric region for Taiwan specifically. The bias is qualitative rather than metric, which is why a geometric probe misses it and a faithfulness-gated verbalization is the sharper instrument.

An anchoring direction exists, but it is not the language direction. The steering turns this into a causal claim and splits it in two. There *is* a linear direction along which "Taiwan" can be pushed into a China-anchored frame: added to an unsteered Mandarin-context vector it raises anchoring from 0 to two-thirds in a frame that never anchors on its own, its removal drives anchoring below baseline, and norm-matched random directions do nothing.

But the direction that does this is Δ_{frame}, estimated from stimulus metadata, not Δ_{lang}, which leaves the non-political frames unmoved. Switching context language therefore does not itself constitute the frame shift; it moves the representation along a frame axis already there. That is also why

the geometric probe of §4.2 was silent, measuring how far the switch moves the vector rather than its component along that axis.

Connection to the English pivot. Our finding is an entity-level, framing-level instance of the broader observation that multilingual models compute over English-aligned representations (Wendler et al., 2024; Schut et al., 2025): for a politically contested entity, "thinking in English" is not neutral. Why English wording rather than simplified orthography carries the effect is left to future work.

Orthography is not stance; provenance is. Because simplified characters are not a proxy for stance (anchoring rides on English vocabulary) while the anchoring itself tracks provenance, audits of Chinese-capable models should measure geopolitical framing directly and evaluate sovereign-AI candidates at the representational level, not only on downstream benchmarks.

6 Limitations

The sharpest concern is that our frame labels come from LLM annotators that may share the priors we measure. Annotators code anchoring from the AV text alone, before the stimulus and its context language are revealed for the drift covariate, and on the validation sample binary-anchoring agreement between the automatic labels and each human (κ=0.45–0.62) is comparable to agreement between the two humans (0.54; Table 7). Moreover, the steering estimand is a per-vector *paired* difference against each vector's own unsteered baseline, so a systematic annotator offset cancels and cannot manufacture the dose-response or the sign reversal.

Our coverage of models and layers is set by the readout. Each NLA is trained for one model at one layer (Fraser-Taliente et al., 2026), and substituting a readout without a reconstructor would forfeit the round-trip check our claims rest on. The released checkpoints cover four models (Qwen2.5-7B, Gemma-3-12B, Gemma-3-27B and Llama-3.3-70B), each at about two-thirds depth, and no Taiwan-adapted model; we use the two in the 7B–12B range so that the contrast is between families rather than scales. We leave further models and layers to future work but state our predictions. Across depth, we expect the gap to be weak in early layers, where the mention's script dominates (in Design B, script effects still match language effects at the mention token even at the depth we study), and near its peak in the English-aligned middle layers (Wendler et al., 2024; Zhao et al., 2024). Across families, since anchoring rides on English vocabulary rather than simplified orthography (§4.5), its direction should hold in any model trained on substantial English text, with magnitude tracking the share of mainland-sourced Chinese data; with scale, testable on Gemma-3-27B, any narrowing should come from Mandarin-context anchoring rising rather than English-context anchoring falling. Among Taiwan-adapted models, Gemma-3-TAIDE-12B (TAIDE, 2025) and Llama-3-Taiwan-70B (Lin and Chen, 2023; Lin, 2024) have no NLA of their own but share architectures with Gemma-3-12B and Llama-3.3-70B, so whether those verbalizers transfer can be checked by round-trip error, whereas Breeze-7B (Hsu et al., 2024) would need a new NLA. To the extent that such adaptation adds Traditional Chinese rather than English data, we expect it to lower Mandarin-context anchoring while leaving English-context anchoring intact, so the gap should persist: localization in Chinese alone may not change the frame a model brings to "Taiwan" for users who write in English.

Some cells are small (GEOPOL-OTHER, n=141; HALLUCINATED, n=133), so those rates are directional, and we do not control for cross-language differences in the mention's surprisal, which multiple templates bound but do not eliminate.

Finally, our claims concern the representation, not generated text. Steering reads out through verbalization, so it establishes a manipulable direction in the representation the verbalizer consumes, not that steering the forward pass changes what the model writes (Turner et al., 2024; Rimsky et al., 2024); round-trip error was not carried through the steered merge, so we treat the largest doses as directional. Likewise, a China-anchored representation need not yield a China-centered statement, since decoding and alignment can suppress it. The findings still matter: output-level audits can miss a disposition the representation encodes, and systems that consume hidden states directly, such as retrieval or classifiers, inherit it regardless. Output-level evidence in fact points the other way, with several models taking more China-aligned positions on Taiwan's sovereignty when queried in Chinese than in English (Ko, 2026). Because our anchoring measures framing more than stance, capturing whether Taiwan is read in relation to China more than which

position on its sovereignty is taken, the two findings need not conflict, but the contrast suggests that representational framing and stated stance can come apart. Linking them within one model requires steering the forward pass and annotating the generated text, or the model's own continuations of our stimuli, under the same codebook, which we leave to future work.

7 Conclusion

We recast a question usually studied through model outputs as a question about the entity's internal representation, answered with a faithfulness-gated verbalization method. Holding propositional content fixed, both Qwen2.5-7B and Gemma-3-12B anchor "Taiwan" to a China-centered geopolitical frame far more often in English than in Mandarin contexts, an effect specific to Taiwan, concentrated at the mention token, and carried by English wording rather than simplified characters.

It is an effect on *which frame* the representation occupies rather than *how far* it moves, which is why a distance-based probe is silent where a frame-level readout is not. Steering then establishes that frame as an encoded, manipulable direction, though the carrier is the frame-metadata direction rather than the language difference, and the demonstration lives in the representation the verbalizer consumes rather than in generated text. The language a user writes in can thus silently change which frame a model brings to a culturally contested entity, giving the sovereign-AI concern a concrete, representation-level signature.

References

Dhruv Agarwal, Mor Naaman, and Aditya Vashistha. 2025. Ai suggestions homogenize writing toward western styles and diminish cultural nuances. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, pages 1–21. ACM.

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. 2024. Refusal in language models is mediated by a single direction. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Arnav Arora, Lucie-Aimée Kaffee, and Isabelle Augenstein. 2023. Probing pre-trained language models for cross-cultural differences in values. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*.

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

R. Harald Baayen, Douglas J. Davidson, and Douglas M. Bates. 2008. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 59(4):390–412.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (COLING-ACL)*.

Dale J. Barr, Roger Levy, Christoph Scheepers, and Harry J. Tily. 2013. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, 68(3):255–278.

Nora Belrose, Zach Furman, Logan Smith, Danny Halawi, Igor Ostrovsky, Lev McKinney, Stella Biderman, and Jacob Steinhardt. 2023. Eliciting latent predictions from transformers with the tuned lens. *arXiv preprint arXiv:2303.08112*.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, et al. 2023. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*.

Yong Cao, Li Zhou, Seolhwa Lee, Laura Cabello, Min Chen, and Daniel Hershcovich. 2023. Assessing cross-cultural alignment between chatgpt and human societies: An empirical study. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*.

Haozhe Chen, Carl Vondrick, and Chengzhi Mao. 2024. SelfIE: Self-interpretation of large language model embeddings. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

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

Esin Durmus, Karina Nguyen, Thomas I. Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, Liane Lovitt, Sam McCandlish, Orowa Sikder, Alex Tamkin, Janel Thamkul, Jared Kaplan, Jack Clark, and Deep Ganguli. 2023. Towards measuring the representation of subjective global opinions in language models. *arXiv preprint arXiv:2306.16388*.

Charles J. Fillmore. 1982. Frame semantics. In The Linguistic Society of Korea, editor, *Linguistics in the Morning Calm*, pages 111–137. Hanshin Publishing Company, Seoul.

Kit Fraser-Taliente, Subhash Kantamneni, Euan Ong, Dan Mossing, Christina Lu, Paul C. Bogdan, Emmanuel Ameisen, James Chen, Dzmitry Kishylau, Adam Pearce, Julius Tarng, Alex Wu, Jeff Wu, Yang Zhang, Daniel M. Ziegler, Evan Hubinger, Joshua Batson, Jack Lindsey, Samuel Zimmerman, and Samuel Marks. 2026. Natural language autoencoders produce unsupervised explanations of LLM activations. *Transformer Circuits Thread*.

Gemma Team. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Asma Ghandeharioun, Avi Caciularu, Adam Pearce, Lucas Dixon, and Mor Geva. 2024. Patchscopes: A unifying framework for inspecting hidden representations of language models. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30).

Hsin-Yi Hsieh, Shih-Cheng Huang, and Richard Tzong-Han Tsai. 2024. TWBias: A benchmark for assessing social bias in traditional Chinese large language models through a Taiwan cultural lens. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 8688–8704, Miami, Florida, USA.

Chan-Jan Hsu, Chang-Le Liu, Feng-Ting Liao, Po-Chun Hsu, Yi-Chang Chen, and Da-Shan Shiu. 2024. Breeze-7B technical report. *arXiv preprint arXiv:2403.02712*.

Ju-Chun Ko. 2026. Bilingual bias in large language models: A Taiwan sovereignty benchmark study. *arXiv preprint arXiv:2602.06371*.

Grgur Kovač, Masataka Sawayama, Rémy Portelas, Cédric Colas, Peter Ford Dominey, and Pierre-Yves Oudeyer. 2023. Large language models as superpositions of cultural perspectives. *arXiv preprint arXiv:2307.07870*.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Ronald W. Langacker. 1987. *Foundations of Cognitive Grammar, Volume 1: Theoretical Prerequisites*. Stanford University Press.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023. Inference-time intervention: Eliciting truthful answers from a language model. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Millicent Li, Alberto Mario Ceballos Arroyo, Giordano Rogers, Naomi Saphra, and Byron C. Wallace. 2025. Do activation verbalization methods convey privileged information? *arXiv preprint arXiv:2509.13316*. Accepted at ICML 2026.

Yen-Ting Lin. 2024. Llama-3-Taiwan-70B-Instruct. https://huggingface.co/yentinglin/Llama-3-Taiwan-70B-Instruct.

Yen-Ting Lin and Yun-Nung Chen. 2023. Taiwan LLM: Bridging the linguistic divide with a culturally aligned language model. *arXiv preprint arXiv:2311.17487*.

Hanjia Lyu, Jiebo Luo, Jian Kang, and Allison Koenecke. 2025. Characterizing bias: Benchmarking large language models in simplified versus traditional Chinese. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*.

Qwen Team. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Aida Ramezani and Yang Xu. 2023. Knowledge of cultural moral norms in large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. 2024. Steering Llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*.

Lisa Schut, Yarin Gal, and Sebastian Farquhar. 2025. Do multilingual LLMs think in English? *arXiv preprint arXiv:2502.15603*.

Mingjie Sun, Xinlei Chen, J. Zico Kolter, and Zhuang Liu. 2024. Massive activations in large language models. In *Conference on Language Modeling (COLM)*.

TAIDE. 2025. Gemma-3-TAIDE-12b-Chat. https://huggingface.co/taide/Gemma-3-TAIDE-12b-Chat.

Zhi-Rui Tam, Ya-Ting Pai, Yen-Wei Lee, Jun-Da Chen, Wei-Min Chu, Sega Cheng, and Hong-Han Shuai. 2024. An improved traditional chinese evaluation suite for foundation model. *arXiv preprint arXiv:2403.01858*.

Yan Tao, Olga Viberg, Ryan S Baker, and René F Kizilcec. 2024. Cultural bias and cultural alignment of large language models. *PNAS Nexus*, 3(9).

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Adam Jermyn, et al. 2024. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. *Transformer Circuits Thread*.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. 2024. Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248*.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Simas Sakenis, Jason Huang, Yaron Singer, and Stuart Shieber. 2020. Causal mediation analysis for interpreting neural NLP: The case of gender bias. *Advances in Neural Information Processing Systems (NeurIPS)*.

Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 2024. Do llamas work in English? on the latent language of multilingual transformers. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*.

Yiran Zhao, Wenxuan Zhang, Guizhen Chen, Kenji Kawaguchi, and Lidong Bing. 2024. How do large language models handle multilingualism? In *Advances in Neural Information Processing Systems (NeurIPS)*.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. Can large language models transform computational social science? *Computational Linguistics*, 50(1).

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. 2023. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*.

A Corpus Construction

Table 2 lists the factors the corpus manipulates and Table 3 its composition. Every sentence is realized in both languages and extracted at both sites, so the sentence-level factors and the extraction site are fully crossed.

Factor	Levels	Role
Context language	en / zh-TW	Manipulated
Stimulus frame	8 frames	Stratifying
Entity	Taiwan / JP, IS	DiD control
Mention script	Hanzi / Latin	Form vs. context
Extraction site	A / B	Measurement locus

Table 2: Factorial structure of the minimal-pair corpus. The eight frames are GEO, POL-INT, POL-DOM, ECON, CUL, HIST, LIFE, TRAV; Site A is the mention-final subtoken and Site B the sentence-final token. Mention script is manipulated only in the Design-B subexperiment.

Block	Composition	Sentences
Core (Taiwan)	$8 \times 12 \times 2$	192
Control (JP, IS)	$2 \times 5 \times 6 \times 2$	120
Design B	$2 \times 6 \times 4$	48
Total		**360**

Table 3: Corpus composition. The 360 sentences yield 720 vectors (two extraction sites each) and, at k=5 verbalizations per vector, 3,600 descriptions *per model* and 7,200 in total.

Tokenizer validation. For both tokenizers we verified that all 360 mentions segment as expected and satisfy the length-matching rule, locating spans by offset mapping (which disagrees with naive subtoken search on 50% of items) and reusing that exact tokenization at extraction.

B Example Minimal Pairs

Table 4 shows representative English-Mandarin minimal pairs, one per frame family. Each pair is a back-translation-verified translation equivalent, mentions the entity exactly once mid-sentence, and establishes the frame in a lead-in of at least ten tokens before naming the entity. Entities never appear as proper-name modifiers (for example *Taiwan Strait*), whose extracted vector would not represent the entity itself.

C Annotation Scheme and Reliability

Table 5 defines the five annotation dimensions; Tables 6 and 7 report reliability and validity respectively.

Beyond the headline Fleiss κ (Table 6), the binary anchoring variable reaches a prevalence-adjusted, bias-adjusted κ (PABAK) of 0.84 to 0.94, against 0.85 for the human pilot. The lower four-way D2 score is a kappa paradox rather than a construct failure: 71% of the annotator disagreements lie on the INGROUP/NEUTRAL boundary, and both labels map to "unanchored" in analysis; the human pilot showed the same disagreement structure (74% on the same boundary), so the LLM annotators hesitate where humans hesitate.

For D5, 60 of 67 disagreements are DRIFTED-versus-ON-TOPIC threshold judgments on a continuous spectrum, with no annotator jumping to HALLUCINATED. Eleven of the 240 triple-coded items have a single unresolved dimension with no two-of-three majority (nine on D2, one each on D1 and D4); these amount to 0.9% of the 1,200 cells in the 240-item reliability sample and are treated as missing data. Reliability and validity were computed with a single shared script and the same ANCHORED $= \{$PRC, GEOPOL-OTHER$\}$ definition for the LLM annotators and the human pilot, so the two are directly comparable.

D Supplementary Statistics and Figures

Table 8 gives the anchoring rate by design frame for Taiwan baseline items, confirming that the geopolitical signal is confined to POL-INT and HIST, and Table 9 cross-tabulates orthography (D4) against anchoring (D2), showing that explicit China-anchoring does not concentrate in simplified-orthography descriptions. Figures 1 to 4 visualize the four main descriptive results of §4.1 to §4.4.

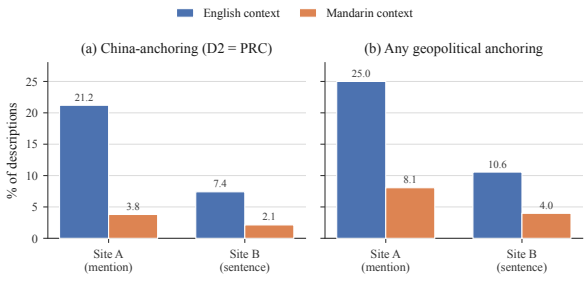

Figure 1: Geopolitical anchoring of "Taiwan" by context language and extraction site. English contexts anchor the entity to a China-centered frame far more often than Mandarin contexts, and the gap is concentrated at the mention token (Site A).

Frame	Sentence pair
GEO	en: From the standpoint of plate tectonics, Taiwan sits right on the boundary between the Eurasian and Philippine Sea plates, which is why earthquakes are so frequent. zh: 從板塊構造的角度來看，台灣正好位於歐亞板塊與菲律賓海板塊的交界，因此地震十分頻繁。
POL-INT	en: In the formal settings of many international organizations, Taiwan's membership status has long been a focal point of negotiation. zh: 在許多國際組織的正式場合裡，台灣的會員資格始終是各方交涉的焦點。
CUL	en: For anyone fond of late-night food hunts, Taiwan's night-market snacks are hard to resist, and the oyster omelet is a must-order. zh: 對許多喜歡深夜覓食的人來說，台灣的夜市小吃是難以抗拒的誘惑，蚵仔煎更是必點。

Table 4: Representative minimal pairs. The international-status frame is de-Sinicized: it discusses standing without naming China, so any China reference in a verbalization is an insertion.

Dim.	Construct	Labels
D1	Exhib. frame	10 classes (8 frames + FOOD, OTHER)
D2	Geopol. anchor	PRC / GEOPOL-OTHER / INGROUP / NEUTRAL (analyzed as anchored vs. not)
D3	Construal	EXPERIENTIAL / ANALYTIC / ACTOR
D4	Orthography	TW-TRAD / PRC-SIMP / NA
D5	Drift (cov.)	ON-TOPIC / DRIFTED / HALLUCINATED

Table 5: The five annotation dimensions. D1 and binary D2 had to clear the pilot reliability threshold before annotation proceeded; D3–D5 are reported but were not gating.

Dim.	Construct	Fleiss κ	Human
D1	Exhibited frame	0.926	0.830
D2	Anchoring (4-way, wt.)	0.609	0.347
D2	Anchoring (binary)	**0.773**	0.459
D3	Construal	0.804	n/a
D4	Orthography	0.870	n/a
D5	Drift	0.619	n/a

Table 6: Inter-annotator agreement over three independent LLM annotators (240 triple-coded items), against the two-annotator human pilot benchmark. The binary anchoring variable, which the analysis uses, is the relevant D2 row.

Dim.	Construct	Human	Mdl.-A	Mdl.-B
D1	Exhibited frame	0.910	0.889	0.854
D2	Anchoring (binary)	0.536	**0.616**	0.448
D3	Construal	0.624	0.590	**0.671**
D4	Orthography	0.922	0.828	0.866
D5	Drift	0.516	0.384	0.449

Table 7: Validity: Cohen's κ between the two human annotators (Human, the ceiling) and between the automatic annotation and each human (Mdl.-A, Mdl.-B), on the stratified 240-item validation sample, which is disjoint from the reliability set of Table 6. Bold marks model-human agreement that meets or exceeds the human ceiling. D1–D4 are valid; only D5 falls short of an already-low ceiling.

Frame	Anchored %	PRC %	n
POL-INT	46.7	17.7	480
HIST	20.8	20.6	480
CUL	7.7	7.7	480
GEO	7.5	7.5	720
POL-DOM	4.8	4.8	480
ECON	4.6	4.4	720
TRAV	4.6	4.6	480
LIFE	4.4	4.4	480

Table 8: Anchoring by design frame (Taiwan baseline). GEO and ECON contribute more items than the other frames.

D4	INGROUP	NEUTRAL	GEOPOL	PRC
TW-TRAD	1045	461	55	48
PRC-SIMP	398	121	10	16

Table 9: Orthography (D4) by anchoring (D2), Taiwan baseline counts. Explicit China-anchoring (PRC) is not concentrated in simplified orthography.

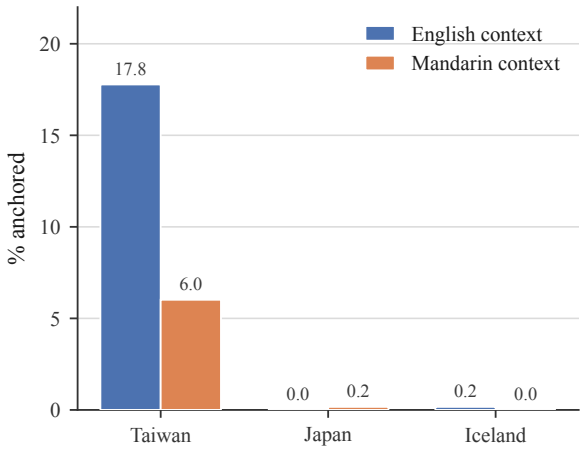

Figure 2: Geopolitical anchoring by entity and context language. Anchoring is near-zero for Japan and Iceland in both languages; only Taiwan shows a large English-over-Mandarin gap.

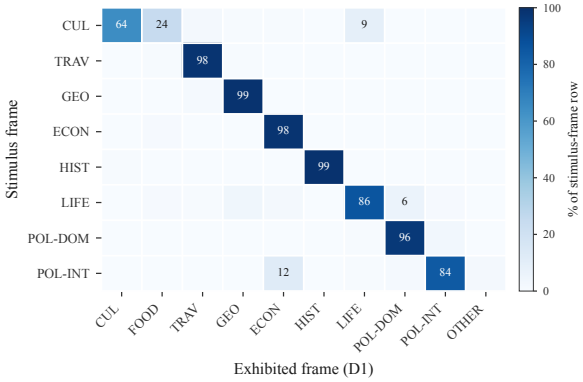

Figure 3: Frame drift matrix for Taiwan (row-normalized %), stimulus frame (rows) to exhibited frame (columns). The strong diagonal shows the pipeline preserves topic; the cross-lingual effect lives in the geopolitical overlay, not in topic identity.

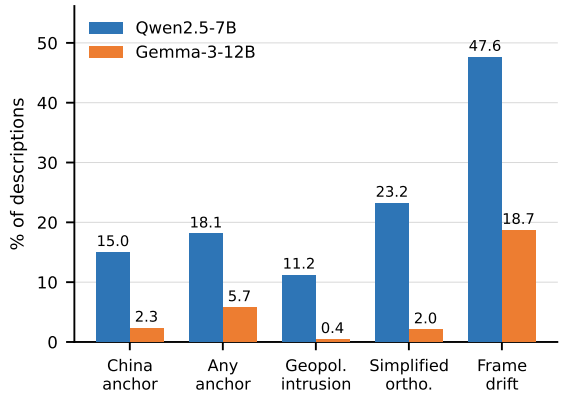

Figure 4: Model comparison on Taiwan baseline descriptions. Qwen exceeds Gemma on every China-leaning and instability index.

E Steering and Analysis Details

Table 10 reports the full dose-response grid behind §4.6 and Figure 5 plots it; the paragraphs below give the test set, the dose scale, and the estimand.

Dose	Hot	Cold	GEO	Δ hot [95% CI]
Gemma-3-12B, Δ_{frame}				
−0.266	0.0	1.1	5.6	−0.22 [−0.39, −0.08]
−0.133	2.8	0.0	0.0	−0.19 [−0.33, −0.06]
−0.065	5.6	0.0	0.0	−0.17 [−0.31, −0.06]
0	22.2	0.0	0.0	n/a
+0.027	27.8	0.0	0.0	+0.06 [−0.06, +0.17]
+0.065	33.3	2.2	0.0	+0.11 [−0.08, +0.33]
+0.133	63.9	20.0	22.2	+0.42 [+0.11, +0.69]
+0.202	72.2	41.1	66.7	+0.50 [+0.17, +0.81]
+0.266	75.0	48.9	66.7	+0.53 [+0.25, +0.78]
Qwen2.5-7B, Δ_{frame}				
−0.296	33.3	0.6	0.0	−0.06 [−0.14, +0.00]
−0.148	30.6	2.2	0.0	0.00 [−0.04, +0.04]
0	34.7	3.0	2.8	n/a
+0.074	31.9	1.6	2.8	+0.01 [−0.06, +0.08]
+0.148	30.6	4.4	2.8	0.00 [−0.06, +0.06]
+0.296	50.0	3.8	13.9	+0.19 [+0.06, +0.33]

Table 10: Anchoring rate (% of descriptions) under Δ_{frame} steering, by relative dose $\|\alpha\delta\|/\|h\|$ and frame group, with the paired per-vector difference from the unsteered condition in the hot frames. Hot = POL-INT, HIST; cold = the five non-political frames; GEO is excluded from its own direction estimate by LOFO. The two models use different dose grids because α is not comparable across them.

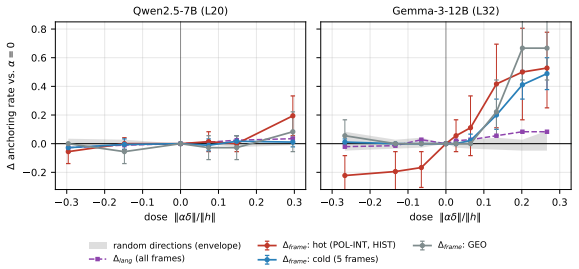

Figure 5: Dose-response of geopolitical anchoring under steering, plotted against relative dose so the two models are comparable. Change is measured per vector against its own unsteered baseline; error bars are 95% bootstrap CIs. The grey band is the envelope of five norm-matched random directions. Δ_{frame} produces a monotone, reversible effect that reaches the held-out GEO frame, whereas Δ_{lang} does not.

Test set and dose. Test vectors are drawn from the Taiwan Site A, Mandarin-context cells only, that is from the *low*-anchoring pole, so the intervention asks whether the representation can be moved into the anchored state that an English context produces on its own. We sample 6 vectors per frame

for Gemma (48 total) and 12 for Qwen (96), draw k=3 verbalizations per steered vector, and report results by frame group: *hot* (POL-INT, HIST, the two frames that carry the descriptive effect), *cold* (CUL, ECON, LIFE, TRAV, POL-DOM), and GEO kept separate because its baseline anchoring rate falls between the two.

Because α is not comparable across models, we report the dose as the relative displacement $\|\alpha\delta\|/\|h\|$: at α=1 this is 0.296 for Qwen but only 0.072 for Gemma, so Gemma's grid is scaled by ≈3.7 to reach a matched maximum dose. The two arms comprise 9,360 (Gemma) and 9,907 (Qwen) newly annotated descriptions.

Estimand. For each test vector we compute the proportion of its k descriptions coded ANCHORED, and take the paired difference against that same vector's unsteered (α=0) proportion; we report the mean paired difference with 95% bootstrap CIs over vectors. Pre-specified reading criteria are monotonic dose-response, sign reversal under $\alpha < 0$, a null for the random directions, and movement in a frame that LOFO excluded from the estimate.

Threats to validity. The three threats addressed by design are output-language confounding, handled by stratifying on description language (§4.1); self-interpretation circularity, handled by parallel replication across two independently trained model families; and the out-of-distribution status of code-switched items, handled by reporting their round-trip error. The language × entity interaction of §3.7 separates because the controls are almost never anchored (two anchored descriptions in 2,400), which is why Taiwan specificity rests on the DiD and the within-Taiwan cluster-robust test rather than on the interaction coefficient.

F Geometric Analysis Details

The variance concentration diagnostic (largest single-dimension share of total variance) is 0.036 for Qwen and 0.468 for Gemma, so Gemma's raw space is dominated by massive-activation dimensions and all geometric quantities are computed after standardizing each dimension to unit variance. The language-displacement magnitude $\|\Delta_{\text{lang}}\|$ is computed per template as the standardized-space distance between the English and Mandarin members of a pair, then averaged by entity and site; it spans entities (P1b) and code-switching cells (P2b). In standardized space Qwen's values are near-identical across entities (≈ 60 to 64), while Gemma's Taiwan value at Site A (58.3) sits below both controls (Japan 70.5, Iceland 64.5), the reversed, control-larger pattern reported in §4.2.

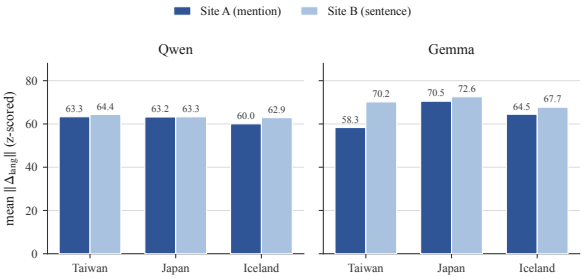

Figure 6: Magnitude of the language-displacement vector $\|\Delta_{\text{lang}}\|$ (standardized space) by entity and site. Taiwan does not move more than the controls under a language switch; for Gemma at Site A it moves less. The annotation-level Taiwan specificity is thus not visible as a larger geometric displacement.