NATURALNESS IS IN THE HEADS: LOCALIZING AND ENHANCING TRANSLATION NATURALNESS IN LLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Modern large language models can generally follow translation instructions; the central challenge is how natural their translations are. To inform quality improvement, mechanistic analysis should explain how models choose among correct translations, not only how they follow translation instructions. Naturalness unfolds over complete sequences and admits multiple valid wordings, so neither a word-level translation readout nor a single reference adequately captures the behavior. We find that the top-ranked 1% of attention heads can causally shift translation naturalness. We localize them using the equivalence-class score difference (Equi-DIFF), which contrasts cardinality-normalized scores over classes of valid multi-token translations and reduces exactly to logit difference in the single-token singleton case. Comparisons against word-level translation and sequence-singleton baselines show that complete-sequence scoring materially changes head rankings, while multi-completion classes improve calibration and reproducibility. Held-out knockout reduces natural-style preference before broad quality degradation, whereas activation steering enhances judged naturalness (69–87% model-level decisive win rates). Across three models, 12 translation directions, and two corpora, decisive naturalness judgments favor steering in every reported setting, providing consistent causal evidence for naturalness heads. Code will be made available after publication.

1 INTRODUCTION

High-quality translation must preserve source meaning and read naturally in the target language. General-purpose large language models (LLMs) can translate through prompting and in-context examples (Zhang et al., 2023; Vilar et al., 2023; Zhu et al., 2024b), while translation-specialized systems use continued pretraining and instruction tuning (Xu et al., 2024a; Alves et al., 2024), preference optimization (Xu et al., 2024b; Zhu et al., 2024a), or iterative refinement (Feng et al., 2024) to improve quality. This progress has narrowed the gap between LLMs and dedicated machine-translation systems, but high aggregate quality does not ensure target-native expression. Models can preserve meaning while retaining source-language syntax, collocations, and idioms. Such "translationese" remains a documented problem in current LLMs, and can be amplified by supervised translation data (Li et al., 2025). The practical question is therefore how a model chooses among correct translations and how that choice can be controlled at inference time while fidelity is evaluated separately.

Output-level evaluation and training interventions reveal the phenomenon, but not the computation that produces it. Multilingual interpretability work finds structured internal language processing, including an English-biased latent space (Wendler et al., 2024) and language-dependent routing in mixture-of-experts models (Bandarkar et al., 2025). Translation-focused studies have further identified components for target-language selection, token alignment, and sentence equivalence (Wang et al., 2024; Zhang et al., 2025; BINBINLIU et al., 2026; Lasnier et al., 2026). These objectives distinguish correct translation from language mismatch or semantic failure. We instead study a finer decision: given two meaning-preserving strategies, which components favor a natural target-language rendering over a literal one?

Causal tracing methods such as activation patching and Patchscopes can localize internal computations (Zhang & Nanda, 2024; Ghandeharioun et al., 2024), while sparse activation interventions can steer

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

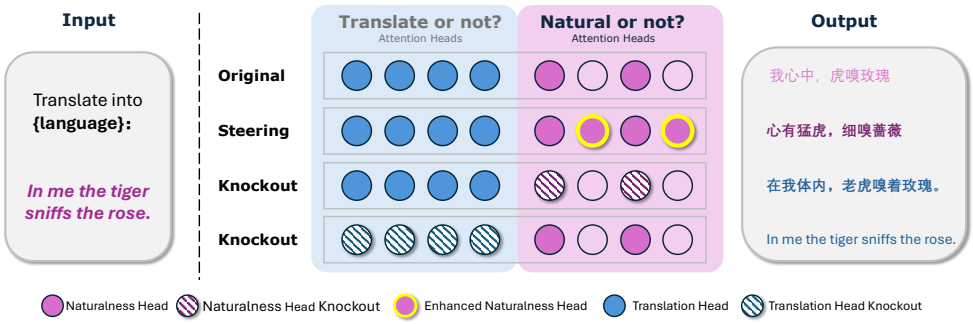

Figure 1: **Translation naturalness as a set-valued sequence behavior.** A source sentence admits multiple valid natural and literal translations, motivating a contrast between completion classes rather than individual strings.

model behavior at inference time (Li et al., 2023; Todd et al., 2024; Rimsky et al., 2024). Together they support a localize–validate–control workflow. However, activation-patching conclusions depend on the behavioral metric (Zhang & Nanda, 2024), and the usual next-token readouts do not directly represent translation naturalness.

Naturalness presents two linked measurement challenges. First, it unfolds over a complete translation rather than a single next token, so a conventional logit difference may capture only the first lexical choice. Second, each source sentence admits multiple valid natural and literal translations, so even a complete-sequence singleton makes one arbitrary wording stand in for an entire behavioral category. As Figure 1 illustrates, the relevant contrast is therefore between sets of complete translations, not between two candidate tokens or two fixed strings.

We introduce the *equivalence-class score difference* (Equi-DIFF) to match this structure. It scores complete translations, aggregates multiple valid realizations with a cardinality-normalized log-mean-exp, and contrasts the natural and literal classes. It generalizes logit difference along the sequence and class axes while retaining the standard single-token case exactly. Prompts differ only in *naturally* versus *literally*; activation patching localizes strategy-conditioning heads, and held-out knockout and activation steering test their causal role.

We make four contributions toward identifying sparse naturalness-modulating components: **(1)** word-level translation and sequence-level naturalness preferentially depend on different head sets; **(2)** across models, direction-specific rankings consistently reuse a substantial within-model core; **(3)** ablating as little as the top-ranked 1% of heads lowers generated natural-style preference across all 72 model–direction–ablation settings; and **(4)** decisive blind naturalness judgments favor translations produced by steering sparse localized components across all 72 model–direction–corpus settings.

2 RELATED WORK

LLM translation has advanced through prompting and translation-specific adaptation (Zhang et al., 2023; Xu et al., 2024a; Alves et al., 2024; Xu et al., 2024b). Yet aggregate quality does not eliminate translationese: supervised translation training can induce source-like syntax and lexical choices (Li et al., 2025), and a graded translationese index captures variation not covered by BLEU or COMET (Liu et al., 2025). Naturalness is therefore distinct from correctness and reference similarity. These output-level studies establish naturalness as a distinct target; we move from improving or evaluating that target to identifying the inference-time components that control it.

Multilingual interpretability has identified an English-biased latent space, language-specific neurons, and language-dependent routing (Wendler et al., 2024; Tang et al., 2024; Bandarkar et al., 2025). Translation-focused studies further localize target-language selection, token alignment, and sentence equivalence (Wang et al., 2024; BINBINLIU et al., 2026; Lasnier et al., 2026). These studies explain whether a model translates into the intended language and preserves meaning, thereby providing the localization foundation for our study. We target a finer decision they leave open: which components choose a target-native rather than source-form literal realization when both preserve meaning.

Activation and path patching localize causal mediators (Vig et al., 2020; Goldowsky-Dill et al., 2023; Ghandeharioun et al., 2024), while sparse activation interventions can steer model behavior (Li et al.,

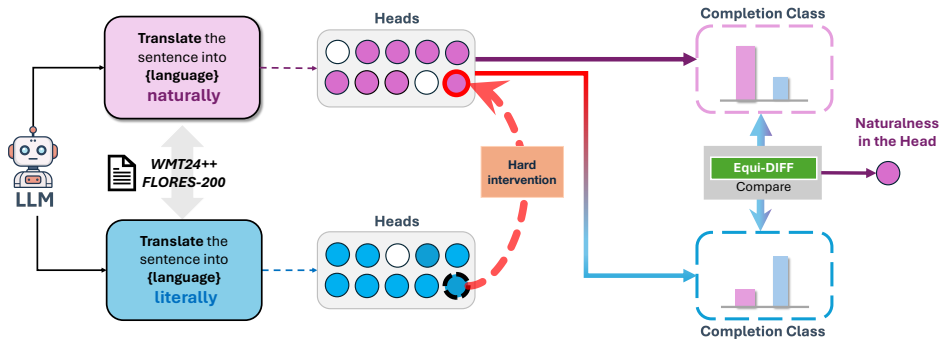

Figure 2: **The Equi-DIFF circuit-analysis pipeline.** Natural and literal prompts induce completion classes whose Equi-DIFF contrast localizes style-sensitive heads; hard interventions on the frozen ranking test their causal contribution to translation naturalness.

2023; Todd et al., 2024; Rimsky et al., 2024). However, localization depends on the behavioral metric (Zhang & Nanda, 2024), and a next-token logit difference cannot represent a sequence-level choice with multiple acceptable realizations. These methods provide our causal toolkit, but not a readout matched to this behavior. Existing work does not jointly provide a class-level, multi-token natural-versus-literal readout, causal localization, and held-out intervention evidence while separately checking semantic adequacy. We connect the three literatures with Equi-DIFF, exact activation/path patching, frozen-ranking knockout, and steering evaluated on both naturalness and adequacy.

3 ATTENTION-HEAD PRELIMINARIES

Attention heads. We use *head* as an operational shorthand for an independently addressable output group immediately before the token mixer's output projection. This boundary is shared by the softmax-attention layers and the Gated DeltaNet layers studied here (Vaswani et al., 2017; Yang et al., 2025), although only the former are multi-head attention in the conventional sense.

For either softmax attention or Gated DeltaNet, let $\mathbf{z}_{\ell,h,t}$ denote output group h immediately before the token-mixer output projection. The intervention boundary is

$$\mathbf{u}_{\ell,t} = W_\ell^O[\mathbf{z}_{\ell,1,t};\ldots;\mathbf{z}_{\ell,H_\ell,t}] = \sum_{h=1}^{H_\ell} W_{\ell,h}^O \mathbf{z}_{\ell,h,t}, \qquad \mathbf{r}_{\ell,t}^+ = \mathbf{r}_{\ell,t} + \Phi_\ell(\mathbf{u}_{\ell,t}). \tag{1}$$

Here $\mathbf{r}_{\ell,t} \in \mathbb{R}^d$ is the residual-stream state entering the token mixer at position t, $\mathbf{r}_{\ell,t}^+$ is the state after its residual update, and Φ_ℓ denotes any configured post-projection transformation, with the identity as the default. The term $W_{\ell,h}^O \mathbf{z}_{\ell,h,t}$ is the independently addressable projected contribution. Localization and steering act before W_ℓ^O; formal knockout removes or replaces the same contribution immediately after projection. Linearity makes these boundaries equivalent at the token-mixer output. Full softmax-attention, gating, GQA, and DeltaNet definitions are in Appendix A.

The resulting localization grids contain 896 output groups for Qwen3.5-4B and 1120 for Qwen3.5-35B-A3B (Team, 2025), and 480 for Gemma-4-26B-A4B-IT (Abd et al., 2026). These indices are model-specific and are therefore not compared directly across models; their layer/type decomposition is provided in Appendix A.

Activation patching. Activation patching uses interchange interventions to measure how replacing an internal activation changes a behavioral readout (Vig et al., 2020; Zhang & Nanda, 2024). We use the path-patching variant (Goldowsky-Dill et al., 2023). Let $\mathcal{A}_{>\ell}$ denote attention outputs downstream of layer ℓ. The core intervention is

$$\mathrm{Patch}_{\ell,h}^{c\to r}(X_r) := \mathrm{do}\big(\mathbf{z}_{\ell,h,t}^r \leftarrow \mathbf{z}_{\ell,h,t}^c \;\forall t,\; \mathbf{z}_{\mathcal{A}_{>\ell},t} \leftarrow \mathbf{z}_{\mathcal{A}_{>\ell},t}^r\big)\,.$$

Thus the selected head is replaced at every scored position, downstream attention is clamped to its reference trajectory, and downstream MLPs recompute normally. Multi-token scoring uses a completion-specific counterfactual cache. We rank units by the resulting readout change, then test the frozen ranking on unseen examples.

4 INTERPRETABILITY ANALYSIS

4.1 DATA AND MODELS

We use the public WMT24++ (Deutsch et al., 2025) and FLORES-200 (Team, 2024) benchmarks. We evaluate translation in both directions between English and German, Spanish, French, Japanese, Russian, and Chinese, giving 12 directions. Our model panel pairs a compact dense anchor, Qwen3.5-4B, with a larger same-family MoE model, Qwen3.5-35B-A3B (Team, 2025), and a cross-family MoE replication, Gemma-4-26B-A4B-IT (Abd et al., 2026). This design tests robustness across model scales and families rather than isolating either factor.

For each model and direction, we select 15 WMT24++ source sentences as the discovery set. We use these examples to construct the completion classes, rank heads, and estimate steering directions. A disjoint held-out knockout set from WMT24++ evaluates interventions on the frozen top-K ranked heads. Steering uses separate held-out examples drawn from WMT24++ and FLORES-200; the WMT24++ steering documents are disjoint from the discovery documents.

Each discovery or knockout example pairs a natural-style reference condition with a literal-style counterfactual condition, as detailed in the supplement. Their prompts differ only by the single-token substitution *naturally*↔*literally*, and each class contains five complete translations generated by the model under study. We use five because the completion-set-size control in Section 4.4 shows that five-member classes produce the most stable localization rankings among the tested sizes. Classes are constructed separately for every model, direction, and prompt condition; the supplement gives the templates, filtering rules, selection criteria, and candidate-generation procedure. The controlled readout analysis first contrasts word-level English–Chinese translation-head localization (Zhang et al., 2025), implemented with a single-token logit difference, with complete-sequence singleton naturalness localization, then replaces the singleton with a five-completion equivalence class.

4.2 EQUI-DIFF METRIC

For a prompt $X \in \{X_r, X_c\}$, X_r denotes the reference prompt containing *naturally*, and X_c denotes the counterfactual prompt containing *literally*. The readout compares a reference completion class $\mathcal{Y}_r$ with a counterfactual completion class $\mathcal{Y}_c$ induced by this one-token prompt contrast. We write the two sets as $\mathcal{Y}_r = \{Y_r^{(i)}\}_{i=1}^{n_r}$ and $\mathcal{Y}_c = \{Y_c^{(j)}\}_{j=1}^{n_c}$, where $n_r = |\mathcal{Y}_r|$ and $n_c = |\mathcal{Y}_c|$ are the respective class cardinalities, and every member is a complete, potentially multi-token translation. For a completion $Y = (y_1, \ldots, y_L)$, $L = |Y|$ denotes its token length. We assign each completion a teacher-forced sequence score

$$s(Y \mid X) = \frac{1}{L^\gamma} \sum_{t=1}^{L} \log p_\theta(y_t \mid X, y_{<t}). \tag{2}$$

Here $\gamma \in [0, 1]$ is the length-normalization exponent: $\gamma = 0$ uses the summed log probability of the complete sequence, whereas $\gamma = 1$ uses its mean token log probability. The aggregation below uses a temperature $\tau > 0$: smaller values concentrate the class score on its highest-scoring completion, while larger values weight class members more evenly.

We aggregate over every member of each completion set and define their contrast directly as

$$\text{Equi-DIFF}(X) = \tau \log \left[\frac{1}{n_r} \sum_{i=1}^{n_r} \exp\left(\frac{s(Y_r^{(i)} \mid X)}{\tau} \right) \right] - \tau \log \left[\frac{1}{n_c} \sum_{j=1}^{n_c} \exp\left(\frac{s(Y_c^{(j)} \mid X)}{\tau} \right) \right]. \tag{3}$$

The comparison unit is therefore a *set of complete translations*, not an individual target sequence. All subsequent experiments use the registered $(\gamma, \tau) = (0, 1)$ setting: it preserves complete-sequence log probability and makes each class term the log of the completion probability averaged under a uniform prior over the enumerated class members. The sensitivity analysis in Section 4.4 further shows that this setting provides stable localization rankings. If the two classes contain one single-token completion a and b, then

$$\text{Equi-DIFF}(X) = \log p(a \mid X) - \log p(b \mid X) = z_a - z_b. \tag{4}$$

The shared softmax normalizer cancels, so Equation 4 recovers conventional logit difference in the single-token singleton case. Equi-DIFF extends this contrast to sequences and completion classes. The $1/n_r$ and $1/n_c$ factors in Equation 3 remove the spurious $\tau \log |\mathcal{Y}|$ offset caused by unequal class sizes. We still use matched cardinalities to maintain comparable sampling precision.

For the patching analysis below, normalization is performed sentence by sentence after patching. We write the normalized effect directly in terms of the three corresponding Equi-DIFF values:

$$E_{\ell,h} = \frac{\text{Equi-DIFF}\big(\text{Patch}_{\ell,h}^{c\to r}(X_r)\big) - \text{Equi-DIFF}(X_r)}{\text{Equi-DIFF}(X_r) - \text{Equi-DIFF}(X_c)}. \tag{5}$$

Here the first term applies the path-patching operator defined above. The denominator normalizes this change by the full unpatched reference-to-counterfactual behavioral gap; sentence-level normalized effects are then averaged.

4.3 HEAD LOCALIZATION

Figure 2 summarizes the analysis. Reference and counterfactual prompts that differ in a single token—*naturally* versus *literally*—define the behavioral contrast, and Equi-DIFF measures that contrast over two sets of complete translations. We patch each addressable attention unit, denote the normalized change in Equi-DIFF by $E_{\ell,h}$, and sort these effects to obtain a ranking. The ranking is frozen before knockout or steering is evaluated; heads with the opposite sign are analyzed separately because they support the contrast in the reverse direction. Model-level aggregate maps and complete direction-specific heatmaps are reported in the supplement (Figures 11 and 10); the frozen ranking is connected to held-out intervention evidence in Finding 3.

4.4 LOCALIZATION COMPARISON AND CONSISTENCY

Two questions determine whether a localized readout captures a reusable mechanism rather than an artifact of task granularity or one sampled wording: **(1)** do word-level translation and sequence-level naturalness depend on different heads; and **(2)** within sequence-level naturalness, is a singleton reference sensitive to which completion is selected, and does a five-completion equivalence class stabilize localization?

To separate these questions, the first analysis contrasts a reproduced word-level translation baseline (Zhang et al., 2025) with complete-sequence naturalness rankings and crosses the frozen rankings with both behavioral endpoints. The second holds the behavioral target and complete-sequence scoring fixed while repeatedly sampling one, three, or five completions per class. Here singleton instability means sensitivity to the chosen natural and literal completions, not optimization or execution noise. Top-K overlap and Jaccard similarity measure agreement at the intervention boundary, while Spearman correlation compares complete rankings. These controls define the frozen five-completion readout subsequently applied without readout retuning across all three models and twelve directions. Because the word-translation and naturalness metrics have different numerical scales, we do not compare their heatmap magnitudes directly; the complete spatial maps are nevertheless provided in Appendix Figure 9.

Finding 1: Word-level translation and sequence-level naturalness preferentially depend on different head sets across models. Cross-ranking knockout reveals a significant readout-by-ranking interaction: word-ranked heads preferentially affect word translation, whereas naturalness-ranked heads preferentially affect held-out sequence-level naturalness. This interaction survives layer/type-matched controls and both zero and mean ablation, supporting preferential functional specialization.

Spatial motivation. The strongest heads also concentrate at different depths: $10/15$ word-level heads lie in L27–L31, whereas $11/15$ formal five-completion naturalness heads lie in L17–L23. The two rankings share only $2/15$ heads (L31.H14 and L18.H9); a fresh complete-sequence singleton likewise shares only $3/15$ heads with the word-translation ranking.

Causal test. Spatial disagreement alone does not establish different functions. We therefore freeze the word-level and five-completion naturalness Top-15 rankings and cross each ranking with both readouts. Each row of Table 1 removes the same 15 heads; each cell is the mean baseline-to-ablation

	Zero ablation		Mean ablation	
Knocked-out heads	Word	Natural.	Word	Natural.
Word heads	**1.33**	0.93	**4.25**	0.37
Naturalness heads	0.62	**3.43**	0.58	**3.11**

Table 1: **Cross-ranking causal summary for the top-ranked 1.7% of heads.** Each cell is the mean baseline-to-ablation loss, normalized by that readout's unablated between-unit SD. Larger values indicate greater causal damage; highlighted cells match each ranking to its own readout.

Top-1 head		Prompt condition		
Type	Unit	Natural	Neutral	Literal
Word	L21.H9	$+.0118$	$+.0130$	$+.0090$
Naturalness	L23.H11	$+.0208$	$+.0202$	$-.0024$

Table 2: **Top-1 translation effects.** ΔNLL; positive values indicate greater dependence than on the matched control. The purple cell marks no positive dependence under the literal prompt.

loss divided by that readout's unablated between-unit SD. Larger values therefore mean greater causal damage on the column behavior.

The crossover provides the core evidence. Under zero ablation, word-ranked heads affect the word readout more ($1.329 > 0.623$), whereas naturalness-ranked heads affect held-out naturalness more ($3.425 > 0.926$). This double dissociation gives $I = (E_{R_W,W} - E_{R_N,W}) + (E_{R_N,N} - E_{R_W,N}) = 3.205$ (95% CI $[2.129, 4.744]$), where $E_{r,e}$ is the standardized effect of ranking r on readout e; mean ablation reproduces it ($I = 6.409$, $[5.362, 7.949]$). Both matched effects also exceed ten layer/type-matched controls, and the interaction remains positive at $K = 5$ and $K = 10$ (Appendix Figure 8). Thus the two rankings are preferentially specialized for different translation behaviors, not merely spatially different; the claim is functional preference, not complete circuit disjointness.

What the two Top-1 localized units do. For the example *"It was raining cats and dogs outside"*, Jacobian-lens (J-Lens) readouts (Gurnee et al., 2026), supported by matched-control interventions, use $\Delta\text{NLL} = (\text{NLL}_h - \text{NLL}_{\text{clean}}) - (\text{NLL}_{h_{\text{ctrl}}} - \text{NLL}_{\text{clean}})$, where h_{ctrl} is the matched control; positive values indicate greater dependence on head h. The top-ranked word-level head L21.H9 consistently supports word-level translation across natural, neutral, and literal prompts (ΔNLL: $+0.0118$, $+0.0130$, $+0.0090$). In contrast, the top-ranked sentence-level head L23.H11 preferentially supports natural translation ($+0.0208$ natural, $+0.0202$ neutral) and shows no positive causal necessity under the literal prompt (-0.0024; Table 2). L21.H9 is therefore more critical to generation continuity during explicit lexical translation, whereas L23.H11 carries sentence-level event and translation semantics and more strongly influences the clean trajectory under neutral and explicitly natural prompts than under the literal prompt. Under neutral prompting, the model defaults to the idiomatic rendering for this particular example. Cross-example localization consistency, these effect profiles, their behavioral consequences, and the J-Lens content jointly support distinct functional priorities. Appendix Figure 10 extends the formal five-completion view across all directions and models.

Finding 2: Within each model, language directions reuse a substantial core of heads. Using the model-specific top 1% of heads ($K = 9$, 12, and 5), consider the two translation orientations separately. Within the six English-to-other-language directions, mean head overlap is 0.548, 0.522, and 0.880 for Qwen3.5-4B, Qwen3.5-35B-A3B, and Gemma, respectively. Among the six other-language-to-English directions, overlap is similarly high at 0.615, 0.722, and 0.813. Thus both orientations show substantial reuse of a compact head set across languages. Pooling all 66 direction pairs, including cross-orientation comparisons, gives mean overlaps of 0.551, 0.581, and 0.667 and circuit-vector cosines of 0.516, 0.520, and 0.722, respectively (Figure 3). All substantially exceed their 1.0–1.1% random-overlap expectations, and cross-orientation overlap remains substantial in all three models (Appendix Table 4). Section 5 tests the frozen rankings on held-out naturalness endpoints.

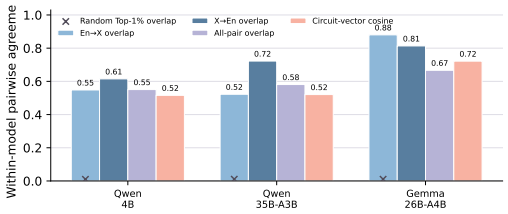

Figure 3: **Within-model head reuse.** The first two bars show within-orientation overlap at model-specific Top-1% budgets. Crosses mark the K/N random-overlap expectation.

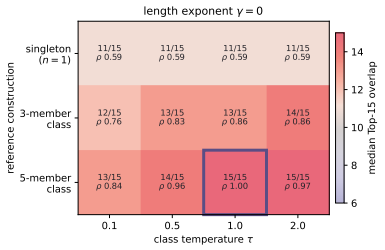

Figure 4: **Completion-choice stability at $\gamma = 0$.** Qwen3.5-4B EN-ZH localization under completion resampling. Cells show median Top-15 overlap and full-ranking Spearman; the outline marks the formal setting.

Metric validation: Multi-completion equivalence classes reduce sensitivity to reference completion choice. Holding the behavioral target and scoring rule fixed, the 25 singleton natural/literal pairs recover a median $11/15$ of the frozen 15-head set (range 6–13), with median full-ranking Spearman 0.59. At $(\gamma, \tau) = (0, 1)$, three-member classes raise the medians to $13/15$ and 0.86, while the registered five-completion estimator reproduces all $15/15$ heads, with full-grid Spearman 0.99934 and maximum normalized-effect deviation 1.21×10^{-4} (Figure 4). This gain in agreement implicates completion choice as the main source of singleton ranking variability and supports the five-completion readout used throughout the study. The $\gamma = 0.5$ and $\gamma = 1$ panels appear in Appendix Figure 7.

5 INTERVENTION RESULTS

Intervention strategies. Knockout ablates prefixes of the frozen head ranking. Zero ablation sets the selected pre-output-projection activations to zero, whereas mean ablation replaces them with their discovery-set mean over reference and counterfactual prompts. We compare the resulting generated class with the fixed counterfactual class using the same five-member Equi-DIFF readout. Generation, sampling, and aggregation details are given in the supplement.

Activation steering uses the discovery-set difference

$$\mathbf{d}_h = \mathbb{E}_i \left[\mathbf{z}^r_{h,i} - \mathbf{z}^c_{h,i} \right] \tag{6}$$

and intervenes during generation using

$$\mathbf{z}_h \leftarrow \mathbf{z}_h + \alpha \mathbf{d}_h. \tag{7}$$

Positive α moves activations toward the natural-prompt direction. Knockout is additionally evaluated at model-normalized Top-1%, 2%, and 3% budgets.

Evaluation protocol. Held-out knockout varies the prefix length K without reranking and evaluates both generated-class Equi-DIFF and translation quality. Steering compares these sparse head sets at $\alpha = 1$ against the same instrumented model without steering ($\alpha = 0$) on WMT24++ and FLORES-200. Naturalness judges see only the paired target-language outputs, whereas adequacy judges also see the source. The evaluator panel contains Qwen3.8-Max, GPT-5.5-0424, Gemini-3.5 Flash, and Claude Opus 4.6; headline multilingual matrices use Gemini-3.5 Flash without pooling judges. In addition to these LLM judges, professional translators manually assessed the naturalness of a sample of model outputs before and after steering. We complement pairwise judgments with BLEU (Papineni et al., 2002), chrF (Popovic, 2015), BERTScore (Zhang et al., 2020), COMET-22 (Rei et al., 2022), and MetricX-24 (Juraska et al., 2024). BLEU, chrF, BERTScore, and COMET are higher-is-better; MetricX is a lower-is-better error score. The supplement specifies prompts, generation, bootstrap scoring, confidence intervals, and coverage reporting.

5.1 MULTILINGUAL KNOCKOUT

Finding 3: Natural-style preference causally relies on sparse ranked heads across models and directions. After ablating the model-specific top 1% of heads ($K = 9$, 12, and 5), both mean and zero ablation lower generated-class Equi-DIFF in all 72 model–direction–ablation settings. Thus

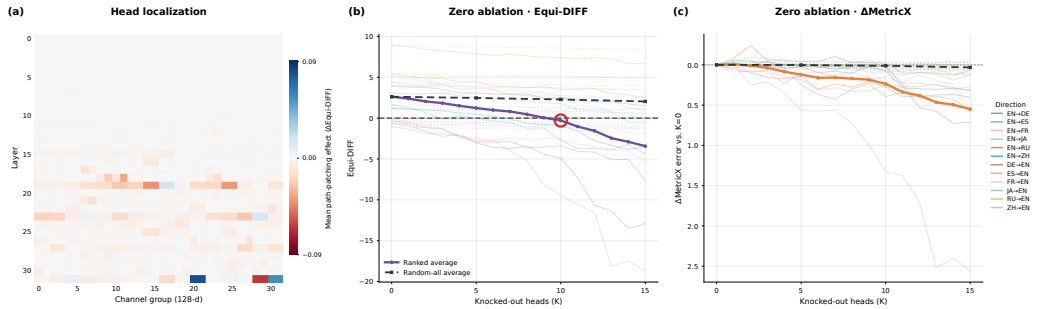

Figure 5: **Qwen3.5-4B localization and held-out knockout.** (a) Twelve-direction mean localization; (b) raw Equi-DIFF and (c) ΔMetricX under zero ablation. The thick purple and orange curves are ranked macro averages; thin colors identify language directions in the right-hand legend, and dashed curves are random-all controls. The red circle in (b) marks the macro average's first point below zero at $K = 10$. The random-all MetricX trajectory uses the shared formal $K = 0$ baseline. The complete analysis is in Appendix Figure 12.

removing approximately 1% of addressable heads reduces natural-style preference in every tested model and direction. Figure 5 links the Qwen3.5-4B head map to its behavioral and translation-quality trajectories; complete three-model results appear in Appendix D.

Representative direction. For the English-to-Chinese held-out direction, mean ablation moves generated-class Equi-DIFF from $+1.191$ at $K = 0$ to -3.524 after removing the top 15 heads (1.7%); zero ablation reaches -3.491. In Figure 5(b), the English-to-Chinese curve crosses zero at seven heads, while the 12-direction macro average crosses below zero at $K = 10$ (red circle). This indicates a shift from the natural toward the literal completion class. At larger K, rising MetricX error shows that increasingly negative Equi-DIFF also coincides with general translation degradation. By contrast, at $K = 15$ the random-all ensemble remains at $+0.687$ and $+0.596$ under mean and zero ablation, respectively. Full results appear in the supplement. **Blind naturalness validation.** Paired judgments compare knockout and $K = 0$ outputs without exposing the source or a reference to Qwen3.8-Max. Across the primary comparisons, knockout outputs lose 0.199 naturalness points on average and win 45.4% of decisive judgments. The clearest reductions occur at 2–3%, where win rates are 42.8–45.9% and mean changes are -0.240 to -0.291.

Cross-direction replication. Across all 12 Qwen3.5-4B directions, both ablation modes reduce Equi-DIFF more for ranked than random-all heads: macro changes are -5.649 versus -0.443 for mean ablation and -6.047 versus -0.581 for zero ablation. MetricX shows the same specificity, with ranked-versus-random error increases of $+0.534$ versus $+0.011$ under mean ablation and $+0.550$ versus $+0.031$ under zero ablation; the ranked increase is larger in all 12 directions. Ranked Equi-DIFF also decreases in all 12 Qwen3.5-35B-A3B and all 12 Gemma directions. Complete curves and baseline definitions appear in the supplement. Together, the uniform Top-1% effect, blind output validation, and ranked-vs-random separation identify sparse causal support for naturalness rather than generic sensitivity to attention ablation.

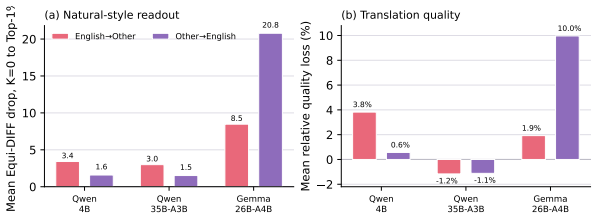

Figure 6: **Model-specific directional knockout sensitivity.** Mean Equi-DIFF and four-metric relative quality loss after removing the model-specific top 1% of heads. Bars average six directions and both ablation modes; negative quality loss denotes an improvement.

Finding 4: Localized heads consistently support natural-reference alignment, with model-specific directional sensitivity. After ablating the model-specific top 1% of heads, generated-class Equi-DIFF decreases in all 72 combinations of model, direction, and ablation mode. At the Top-15 endpoint, knockout also worsens natural-class chrF, BERTScore, and COMET, as well as official-reference MetricX, in every case. In Qwen3.5-35B-A3B, knockout moves generations closer to the counterfactual class: similarity to that class improves in 7/12 directions by BLEU and 6/12 by

COMET under each mode. Together with the consistent decline in natural-class alignment, this directly links the localized heads to the model's natural-versus-literal preference.

Directional sensitivity forms a reproducible model signature. For the top 1% of heads, both Qwen models show larger Equi-DIFF losses for English-to-other-language directions (3.41 versus 1.59 and 2.99 versus 1.51), whereas Gemma is more sensitive for other-language-to-English directions (20.78 versus 8.47). All six model-and-mode paired-language 95% bootstrap intervals remain on one side of zero, so the orientation difference persists under resampling. Quality change remains between a 1.2% improvement and a 10.0% loss at this sparse intervention level (Figure 6).

5.2 MULTILINGUAL STEERING

Finding 5: Sparse localized components provide a reusable cross-model control handle for translation naturalness. Across all 72 model–direction–corpus settings, decisive blind naturalness judgments favor steering. Model-level decisive win rates average 69.25% for Qwen3.5-4B, 82.97% for Qwen3.5-35B-A3B, and 86.63% for Gemma (Table 3). The same table jointly reports paired adequacy and macro-averaged MetricX and COMET changes.

	Qwen3.5-4B				Qwen3.5-35B-A3B				Gemma-4-26B-A4B-IT			
	WMT24++		FLORES-200		WMT24++		FLORES-200		WMT24++		FLORES-200	
Direction	**Nat.**	Adeq.	**Nat.**	Adeq.	**Nat.**	Adeq.	**Nat.**	Adeq.	**Nat.**	Adeq.	**Nat.**	Adeq.
En→De	**57.4**	43.2	**59.0**	42.8	**81.1**	26.8	**82.1**	21.4	**76.8**	59.0	**86.2**	74.0
En→Es	**59.3**	38.0	**69.7**	40.6	**82.8**	25.6	**83.0**	17.6	**88.3**	75.4	**94.5**	83.6
En→Fr	**62.0**	42.1	**64.2**	36.6	**70.3**	38.6	**75.5**	42.5	**80.7**	63.2	**87.0**	76.7
En→Ja	**60.1**	47.8	**60.5**	41.2	**81.5**	29.7	**76.0**	24.6	**88.2**	100.0	**83.3**	–
En→Ru	**60.2**	43.0	**58.5**	33.8	**83.8**	27.8	**81.9**	23.2	**88.6**	73.9	**82.9**	70.0
En→Zh	**72.5**	32.8	**69.1**	30.1	**88.0**	17.8	**82.7**	17.4	**91.5**	73.7	**83.8**	42.9
De→En	**79.7**	25.5	**79.2**	23.2	**84.8**	24.6	**85.1**	18.8	**88.0**	78.2	**86.5**	78.9
Es→En	**76.6**	24.8	**79.7**	22.2	**87.9**	17.1	**81.6**	21.3	**89.7**	65.5	**88.5**	73.8
Fr→En	**77.5**	19.2	**80.4**	20.6	**86.1**	23.5	**84.6**	14.5	**86.9**	68.3	**85.3**	66.0
Ja→En	**71.5**	37.4	**69.2**	37.0	**89.7**	14.7	**85.4**	11.9	**83.3**	54.9	**88.0**	60.9
Ru→En	**75.2**	27.3	**80.0**	21.7	**91.2**	9.3	**89.7**	7.8	**88.2**	74.2	**89.2**	71.5
Zh→En	**71.1**	29.9	**69.6**	34.1	**81.0**	21.0	**75.8**	15.8	**86.7**	66.0	**86.8**	71.8
ΔMetricX | ΔCOMET	+.067	+.0005	+.025	−.0006	+.124	−.0015	+.027	−.0016	+.299	+.0047	+.253	−.0005

Table 3: **Core multilingual steering results.** Decisive naturalness/adequacy win rates (%); rates above 50% favor steering, and green columns mark naturalness. The final row gives 12-direction macro ΔMetricX and ΔCOMET; positive deltas favor steering.

Finding 6: Consistent naturalness gains across evaluators from moderate steering of sparse heads. On fixed Qwen3.5-4B English-to-Chinese generations at $\alpha = 1$, all four evaluators assign higher naturalness after steering: ΔNaturalness is $+0.527$ (Qwen3.8-Max), $+0.451$ (GPT-5.5), $+0.618$ (Gemini-3.5 Flash), and $+0.343$ (Claude Opus 4.6). All four sentence-cluster 95% intervals are positive, with a cross-judge mean of $+0.485 \pm 0.117$ (mean ± between-evaluator SD). In the independent dose sweep, increasing α from 0.5 to 1 raises GPT-5.5's gain from $+0.231$ to $+0.451$ with 15 heads and from $+0.253$ to $+0.518$ with 20 heads; Gemini-3.5 Flash rises from $+0.347$ to $+0.618$ and from $+0.360$ to $+0.683$, respectively. All eight 95% intervals are positive. The near doubling with α, compared with the smaller difference between 15 and 20 heads, identifies steering strength as the dominant factor in this local sweep. Gemma tie rates, valid-output coverage, and illustrative held-out cases are reported in Appendix E (Tables 5 and 6).

6 CONCLUSION

Translation naturalness is a sequence-level choice among valid expressions. Equi-DIFF models this structure while retaining single-token logit difference as a special case. Sequence-level localization differs from word-level translation localization, and equivalence classes improve stability over singleton targets. Held-out knockout lowers natural-style preference in all 72 model, direction, and ablation settings, while blind judgments favor steering in all 72 model, direction, and corpus settings. The rankings reuse a substantial within-model core, yet directional sensitivity remains model-specific. Together, these results identify sparse causal components and a dose-responsive control handle for target-native translation.

AI USE STATEMENT

Generative AI tools assisted code development, experiment documentation, and early manuscript organization. The authors reviewed all AI-assisted work and remain responsible for the final text, claims, code, and artifacts.

REPRODUCIBILITY STATEMENT

Run manifests record model and data identities, prompts, numerical precision, random seeds, generation settings, intervention definitions, and upstream artifact hashes. The supplement provides dataset construction details, frozen method-validation results, evaluation protocols, full per-direction tables, and artifact provenance.

REFERENCES

Sherif El Abd, Vaibhav Aggarwal, Robin Algayres, Alek Andreev, Olivier Bachem, Ian Ballantyne, Cormac Brick, Victor Carbune, Michelle Casbon, Mayank Chaturvedi, Aditya Chawla, Victor Cotruta, Alice Coucke, Phil Culliton, Robert Dadashi, Lucas Dixon, Mohamed Elhawaty, Utku Evci, Clément Farabet, Johan Ferret, Filippo Galgani, Sertan Girgin, Jean-Bastien Grill, Maarten Grootendorst, Jiaxian Guo, Cassidy Hardin, Yanzhang He, Steven M. Hernandez, Omri Homburger, Léonard Hussenot, Ju-yeong Ji, Armand Joulin, Aishwarya Kamath, Parnian Kassraie, Olivier Lacombe, Preethi Lahoti, Gaël Liu, Gus Martins, Luciano Martins, Tatiana Matejovicova, Ramona Merhej, Nikola Momchev, Sneha Mondal, Ryan Mullins, Sindhu Raghuram Panyam, Shreya Pathak, Sarah Perrin, André Susano Pinto, Etienne Pot, Angéline Pouget, Alexandre Ramé, Sabela Ramos, Douglas Reid, David Rim, Morgane Rivière, Karsten Roth, Louis Rouillard, Omar Sanseviero, Pier Giuseppe Sessa, Shane Settle, Danila Sinopalnikov, Sara Smoot, Piotr Stanczyk, Andreas Steiner, Lawrence Stewart, Ilya O. Tolstikhin, Michael Tschannen, Anton Tsitsulin, Nino Vieillard, Renjie Wu, Pingmei Xu, Haichuan Yang, Edouard Yvinec, Biao Zhang, Li Zhang, Joe Zou, Nicolas Aagnes, Abdelrahman Abdelhamed, Jakub Adámek, Shivani Agrawal, Shubham Agrawal, Ibrahim Alabdulmohsin, Jean-Baptiste Alayrac, Uri Alon, Chandramouli Amarnath, Ankesh Anand, Chrysovalantis Anastasiou, Setareh Ariafar, François-Xavier Aubet, Kyriakos Axiotis, Federico Barbero, Joelle K. Barral, Alexei Bendebury, Urs Bergmann, Stanley Bileschi, Kat Black, Mathieu Blondel, Sebastian Borgeaud, and Arthur Brazinskas. Gemma 4 technical report. *CoRR*, abs/2607.02770, 2026. doi: 10.48550/ARXIV.2607.02770. URL https://doi.org/10.48550/arXiv.2607.02770.

Joshua Ainslie, James Lee-Thorp, Michiel de Jong, Yury Zemlyanskiy, Federico Lebrón, and Sumit Sanghai. GQA: training generalized multi-query transformer models from multi-head checkpoints. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pp. 4895–4901. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.EMNLP-MAIN.298. URL https://doi.org/10.18653/v1/2023.emnlp-main.298.

Duarte M. Alves, José Pombal, Nuno Miguel Guerreiro, Pedro Henrique Martins, João Alves, M. Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, Pierre Colombo, José G. C. de Souza, and André F. T. Martins. Tower: An open multilingual large language model for translation-related tasks. *CoRR*, abs/2402.17733, 2024. doi: 10.48550/ARXIV.2402.17733. URL https://doi.org/10.48550/arXiv.2402.17733.

Lucas Bandarkar, Chenyuan Yang, Mohsen Fayyaz, Junlin Hu, and Nanyun (Violet) Peng. Multilingual routing in mixture-of-experts. *CoRR*, abs/2510.04694, 2025. doi: 10.48550/ARXIV.2510.04694. URL https://doi.org/10.48550/arXiv.2510.04694.

BINBINLIU, Wenhan Han, Feng Chen, Yifan Zhang, Ping Guo, Haobin Lin, Bingni Zhang, Taifeng Wang, and Yin Zheng. Token alignment heads: Unveiling attention's role in LLM multilingual translation. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=q8fTgw8e5E.

Daniel Deutsch, Eleftheria Briakou, Isaac Rayburn Caswell, Mara Finkelstein, Rebecca Galor, Juraj Juraska, Geza Kovacs, Alison Lui, Ricardo Rei, Jason Riesa, Shruti Rijhwani, Parker Riley,

Elizabeth Salesky, Firas Trabelsi, Stephanie Winkler, Biao Zhang, and Markus Freitag. WMT24++: expanding the language coverage of WMT24 to 55 languages & dialects. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, volume ACL 2025 of *Findings of ACL*, pp. 12257–12284. Association for Computational Linguistics, 2025. doi: 10.18653/V1/2025.FINDINGS-ACL.634. URL https://doi.org/10.18653/v1/2025.findings-acl.634.

Zhaopeng Feng, Ruizhe Chen, Yan Zhang, Zijie Meng, and Zuozhu Liu. Ladder: A model-agnostic framework boosting llm-based machine translation to the next level. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pp. 15377–15393. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.EMNLP-MAIN.860. URL https://doi.org/10.18653/v1/2024.emnlp-main.860.

Asma Ghandeharioun, Avi Caciularu, Adam Pearce, Lucas Dixon, and Mor Geva. Patchscopes: A unifying framework for inspecting hidden representations of language models. In Ruslan Salakhutdinov, Zico Kolter, Katherine A. Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, volume 235 of *Proceedings of Machine Learning Research*, pp. 15466–15490. PMLR / OpenReview.net, 2024. URL https://proceedings.mlr.press/v235/ghandeharioun24a.html.

Nicholas Goldowsky-Dill, Chris MacLeod, Lucas Sato, and Aryaman Arora. Localizing model behavior with path patching. *CoRR*, abs/2304.05969, 2023. doi: 10.48550/ARXIV.2304.05969. URL https://doi.org/10.48550/arXiv.2304.05969.

Wes Gurnee, Nicholas Sofroniew, Adam Pearce, Mateusz Piotrowski, Isaac Kauvar, Runjin Chen, Anna Soligo, Paul C. Bogdan, Euan Ong, Rowan Wang, Ben Thompson, David Abrahams, Subhash Kantamneni, Emmanuel Ameisen, Joshua Batson, and Jack Lindsey. Verbalizable representations form a global workspace in language models. *CoRR*, abs/2607.15495, 2026. doi: 10.48550/ARXIV.2607.15495. URL https://doi.org/10.48550/arXiv.2607.15495.

Juraj Juraska, Daniel Deutsch, Mara Finkelstein, and Markus Freitag. Metricx-24: The google submission to the WMT 2024 metrics shared task. In Barry Haddow, Tom Kocmi, Philipp Koehn, and Christof Monz (eds.), *Proceedings of the Ninth Conference on Machine Translation, WMT 2024, Miami, FL, USA, November 15-16, 2024*, pp. 492–504. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.WMT-1.35. URL https://doi.org/10.18653/v1/2024.wmt-1.35.

Théo Lasnier, Armel Randy Zebaze, Djamé Seddah, Rachel Bawden, and Benoît Sagot. Translation heads: Disentangling meaning from language in LLM-based machine translation. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=JKSDWukEgN.

Kenneth Li, Oam Patel, Fernanda B. Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/81b8390039b7302c909cb769f8b6cd93-Abstract-Conference.html.

Yafu Li, Ronghao Zhang, Zhilin Wang, Huajian Zhang, Leyang Cui, Yongjing Yin, Tong Xiao, and Yue Zhang. Lost in literalism: How supervised training shapes translationese in llms. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pp. 12875–12894. Association for Computational Linguistics, 2025. doi: 10.18653/V1/2025.ACL-LONG.630. URL https://doi.org/10.18653/v1/2025.acl-long.630.

Yikang Liu, Wanyang Zhang, Yiming Wang, Jialong Tang, Pei Zhang, Baosong Yang, Fei Huang, Rui Wang, and Hai Hu. Translationese-index: Using likelihood ratios for graded and generalizable measurement of translationese. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, EMNLP 2025, Suzhou, China, November 4-9, 2025*, pp. 12521–12538. Association for Computational Linguistics, 2025. doi: 10.18653/V1/2025.EMNLP-MAIN.633. URL https://doi.org/10.18653/v1/2025.emnlp-main.633.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*, pp. 311–318. ACL, 2002. doi: 10.3115/1073083.1073135. URL https://aclanthology.org/P02-1040/.

Maja Popovic. chrf: character n-gram f-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation, WMT@EMNLP 2015, 17-18 September 2015, Lisbon, Portugal*, pp. 392–395. The Association for Computer Linguistics, 2015. doi: 10.18653/V1/W15-3049. URL https://doi.org/10.18653/v1/w15-3049.

Zihan Qiu, Zekun Wang, Bo Zheng, Zeyu Huang, Kaiyue Wen, Songlin Yang, Rui Men, Le Yu, Fei Huang, Suozhi Huang, Dayiheng Liu, Jingren Zhou, and Junyang Lin. Gated attention for large language models: Non-linearity, sparsity, and attention-sink-free. In Danielle Belgrave, Cheng Zhang, Laura N. Montoya, Hsuan-Tien Lin, Razvan Pascanu, Piotr Koniusz, Marzyeh Ghassemi, Nancy Chen, Iván Vladimir Meza Ruíz, and Arturo Loaiza-Bonilla (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*, 2025. URL http://papers.nips.cc/paper_files/paper/2025/hash/904e89bb4e632e75fb47f093b620b257-Abstract-Conference.html.

Ricardo Rei, José G. C. de Souza, Duarte M. Alves, Chrysoula Zerva, Ana C. Farinha, Taisiya Glushkova, Alon Lavie, Luísa Coheur, and André F. T. Martins. COMET-22: unbabel-ist 2022 submission for the metrics shared task. In Philipp Koehn, Loïc Barrault, Ondrej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno-Yepes, Tom Kocmi, André F. T. Martins, Makoto Morishita, Christof Monz, Masaaki Nagata, Toshiaki Nakazawa, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Marco Turchi, and Marcos Zampieri (eds.), *Proceedings of the Seventh Conference on Machine Translation, WMT 2022, Abu Dhabi, United Arab Emirates (Hybrid), December 7-8, 2022*, pp. 578–585. Association for Computational Linguistics, 2022. doi: 10.18653/V1/2022.WMT-1.52. URL https://doi.org/10.18653/v1/2022.wmt-1.52.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. Steering llama 2 via contrastive activation addition. 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), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pp. 15504–15522. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.ACL-LONG.828. URL https://doi.org/10.18653/v1/2024.acl-long.828.

Tianyi Tang, Wenyang Luo, Haoyang Huang, Dongdong Zhang, Xiaolei Wang, Xin Zhao, Furu Wei, and Ji-Rong Wen. Language-specific neurons: The key to multilingual capabilities in large language models. 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), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pp. 5701–5715. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.ACL-LONG.309. URL https://doi.org/10.18653/v1/2024.acl-long.309.

NLLB Team. Scaling neural machine translation to 200 languages. *Nat.*, 630(8018):841–846, 2024. doi: 10.1038/S41586-024-07335-X. URL https://doi.org/10.1038/s41586-024-07335-x.

Qwen Team. Qwen3 technical report. *CoRR*, abs/2505.09388, 2025. doi: 10.48550/ARXIV.2505.09388. URL https://doi.org/10.48550/arXiv.2505.09388.

Eric Todd, Millicent L. Li, Arnab Sen Sharma, Aaron Mueller, Byron C. Wallace, and David Bau. Function vectors in large language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=AwyxtyMwaG.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 5998–6008, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart M. Shieber. Causal mediation analysis for interpreting neural NLP: the case of gender bias. *CoRR*, abs/2004.12265, 2020. URL https://arxiv.org/abs/2004.12265.

David Vilar, Markus Freitag, Colin Cherry, Jiaming Luo, Viresh Ratnakar, and George F. Foster. Prompting palm for translation: Assessing strategies and performance. In Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pp. 15406–15427. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.ACL-LONG.859. URL https://doi.org/10.18653/v1/2023.acl-long.859.

Weichuan Wang, Zhaoyi Li, Defu Lian, Chen Ma, Linqi Song, and Ying Wei. Mitigating the language mismatch and repetition issues in llm-based machine translation via model editing. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pp. 15681–15700. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.EMNLP-MAIN.879. URL https://doi.org/10.18653/v1/2024.emnlp-main.879.

Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. Do llamas work in english? on the latent language of multilingual transformers. 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), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pp. 15366–15394. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.ACL-LONG.820. URL https://doi.org/10.18653/v1/2024.acl-long.820.

Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. A paradigm shift in machine translation: Boosting translation performance of large language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024a. URL https://openreview.net/forum?id=farT6XXntP.

Haoran Xu, Amr Sharaf, Yunmo Chen, Weiting Tan, Lingfeng Shen, Benjamin Van Durme, Kenton Murray, and Young Jin Kim. Contrastive preference optimization: Pushing the boundaries of LLM performance in machine translation. In Ruslan Salakhutdinov, Zico Kolter, Katherine A. Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, volume 235 of *Proceedings of Machine Learning Research*, pp. 55204–55224. PMLR / OpenReview.net, 2024b. URL https://proceedings.mlr.press/v235/xu24t.html.

Songlin Yang, Jan Kautz, and Ali Hatamizadeh. Gated delta networks: Improving mamba2 with delta rule. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=r8H7xhYPwz.

Biao Zhang, Barry Haddow, and Alexandra Birch. Prompting large language model for machine translation: A case study. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Pro-*

ceedings of Machine Learning Research, pp. 41092–41110. PMLR, 2023. URL https://proceedings.mlr.press/v202/zhang23m.html.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=Hf17y6u9BC.

Hongbin Zhang, Kehai Chen, Xuefeng Bai, Xiucheng Li, Yang Xiang, and Min Zhang. Exploring the translation mechanism of large language models. In Danielle Belgrave, Cheng Zhang, Laura N. Montoya, Hsuan-Tien Lin, Razvan Pascanu, Piotr Koniusz, Marzyeh Ghassemi, Nancy Chen, Iván Vladimir Meza Ruíz, and Arturo Loaiza-Bonilla (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*, 2025. URL http://papers.nips.cc/paper_files/paper/2025/hash/99367869dc65679f7bc243b45a23a92f-Abstract-Conference.html.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=SkeHuCVFDr.

Dawei Zhu, Sony Trenous, Xiaoyu Shen, Dietrich Klakow, Bill Byrne, and Eva Hasler. A preference-driven paradigm for enhanced translation with large language models. In Kevin Duh, Helena Gómez-Adorno, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pp. 3385–3403. Association for Computational Linguistics, 2024a. doi: 10.18653/V1/2024.NAACL-LONG.186. URL https://doi.org/10.18653/v1/2024.naacl-long.186.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. Multilingual machine translation with large language models: Empirical results and analysis. In Kevin Duh, Helena Gómez-Adorno, and Steven Bethard (eds.), *Findings of the Association for Computational Linguistics: NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, volume NAACL 2024 of *Findings of ACL*, pp. 2765–2781. Association for Computational Linguistics, 2024b. doi: 10.18653/V1/2024.FINDINGS-NAACL.176. URL https://doi.org/10.18653/v1/2024.findings-naacl.176.

A Attention and Intervention Details

Let $\mathbf{r}_{\ell,t} \in \mathbb{R}^d$ be the residual-stream vector entering the token mixer and $\widetilde{\mathbf{r}}_{\ell,t} = \mathrm{Norm}_\ell(\mathbf{r}_{\ell,t})$. In a softmax-attention layer (Vaswani et al., 2017), the projections and head output are

$$\begin{aligned} \mathbf{q}_{\ell,h,t} &= W^Q_{\ell,h}\widetilde{\mathbf{r}}_{\ell,t}, & \mathbf{k}_{\ell,j,t} &= W^K_{\ell,j}\widetilde{\mathbf{r}}_{\ell,t}, & \mathbf{v}_{\ell,j,t} &= W^V_{\ell,j}\widetilde{\mathbf{r}}_{\ell,t}, \\ a_{\ell,h,t,s} &= \underset{s\in\mathcal{C}_{\ell,t}}{\mathrm{softmax}}\left(\frac{\mathbf{q}^\top_{\ell,h,t}\mathbf{k}_{\ell,\rho_\ell(h),s}}{\sqrt{d_k}}\right), & \mathbf{o}_{\ell,h,t} &= \sum_{s\in\mathcal{C}_{\ell,t}} a_{\ell,h,t,s}\mathbf{v}_{\ell,\rho_\ell(h),s}. \end{aligned} \tag{8}$$

The map ρ_ℓ assigns query heads to key/value heads. It is the identity for standard multi-head attention; in GQA, several query heads share one key/value head (Ainslie et al., 2023). The context set $\mathcal{C}_{\ell,t}$ is either the full causal prefix or a bounded sliding window. The released models additionally apply their configured normalization, rotary encoding, and logit transformations upstream of the intervention.

Qwen3.5 gates the softmax-attention output (Qiu et al., 2025):

$$\mathbf{z}_{\ell,h,t} = \sigma\big(W^G_{\ell,h}\widetilde{\mathbf{r}}_{\ell,t}\big) \odot \mathbf{o}_{\ell,h,t}. \tag{9}$$

Gemma's GQA layers instead use $\mathbf{z}_{\ell,h,t} = \mathbf{o}_{\ell,h,t}$. Qwen3.5 also contains Gated DeltaNet layers (Yang et al., 2025), whose recurrent core is

$$\mathbf{S}_t = \alpha_t\mathbf{S}_{t-1}\left(I - \beta_t\mathbf{k}_t\mathbf{k}_t^\top\right) + \beta_t\mathbf{v}_t\mathbf{k}_t^\top, \qquad \mathbf{o}_t = \mathbf{S}_t\mathbf{q}_t. \tag{10}$$

Here α_t is the state-decay gate and β_t the write-strength gate. We denote the module's value/output group immediately before `out_proj` by $\mathbf{z}_{\ell,h,t}$ without interpreting it as a softmax-attention head. No intervention modifies attention weights or query/key groups.

The released configurations instantiate this boundary differently (Team, 2025; Abd et al., 2026). Qwen3.5-4B has 8 gated-GQA layers with 16 output heads and 24 Gated DeltaNet layers with 32 output groups; Qwen3.5-35B-A3B has 10 gated-GQA and 30 Gated DeltaNet layers. Gemma-4-26B-A4B-IT has 25 sliding-GQA and 5 global-GQA layers, each with 16 output heads. The corresponding value widths, key/value-head counts, and intervention implementation details are recorded in the released run manifests.

B Experimental Protocol

Prompt contrast and completion classes. For every source sentence, the reference and counterfactual prompts differ by the single-token substitution *naturally↔literally*. Each accepted unit has five reference and five counterfactual completions generated by the model being analyzed. We regenerate these classes for every model, language, and direction; no formal result reuses another model's or direction's completions. Mechanical target-language filters reject prompt echoes, unfinished reasoning traces, wrong-language text, and malformed outputs. A unit is retained when the reference prompt favors the reference class, the counterfactual prompt favors the counterfactual class, and the separation between those two prompt-conditioned scores is at least 1.0. The formal construction samples 100 candidates per prompt before filtering, deduplicates the valid outputs, removes completions shared across the two classes, and retains the five candidates with the highest teacher-forced sequence scores under the class-generating prompt. Thus, the fixed $\mathcal{Y}_r$ and $\mathcal{Y}_c$ are model-generated, scored classes rather than random subsets of the later knockout generations.

Cardinality normalization. The normalized score can be written as $\tau\,\mathrm{logsumexp}(s/\tau) - \tau\log|\mathcal{Y}|$. The second term removes the deterministic benefit of enumerating more surface forms. For example, if all candidates have score q, raw log-sum-exp assigns $q + \tau\log 20$ to a 20-member pool and $q + \tau\log 5$ to a five-member pool. The resulting $\tau\log 4$ difference is entirely due to pool size; normalized scores assign both pools q and are unchanged if every member is duplicated.

The localization contrast and the three steering prompts use the following instruction lines; source and target-language slots are filled without changing the surrounding chat template:

```
Neutral:  Translate the following sentence into {TARGET}:
Natural:  Translate the following sentence naturally into
```

```
{TARGET}:
Literal: Translate the following sentence literally into
{TARGET}:
Sentence: {SOURCE}
Translation:
```

Only the natural/reference and literal/counterfactual prompts define the AP pair. The neutral prompt is an instruction-transfer condition used during steering, not a third localization class. For translation, "counterfactual" refers to the literal style instruction rather than an incorrect semantic answer.

Data roles. For each formal combination of model and translation direction, the *discovery set* contains 15 WMT24++ source sentences and is used to construct reference classes, rank heads, and estimate steering directions. A disjoint WMT24++ split of 15 sentences serves as the *held-out knockout set* and is used only for knockout evaluation after the ranking is frozen. A second *held-out steering set* is drawn from the frozen WMT24++ and FLORES-200 test sets and is used only for generated-output steering evaluation. For every model and direction, we select 400 WMT24++ sentences and 300 FLORES-200 sentences. WMT24++ steering documents are disjoint from discovery documents. Thus discovery, knockout, and steering examples have separate experimental roles.

Models and intervention scope. The primary models are Qwen3.5-4B, Qwen3.5-35B-A3B, and Gemma-4-26B-A4B-IT. The frozen Top-15 sets comprise 1.7% of Qwen3.5-4B's 896 addressable attention units, 1.3% of Qwen3.5-35B-A3B's 1,120 units, and 3.1% of Gemma's 480 units. Localization, knockout, and steering cover all 12 directions between English and German, Spanish, French, Japanese, Russian, and Chinese for all three models.

Interventions. Activation patching (called path patching in the implementation artifacts) ranks pre-output-projection attention units by the change in Equi-DIFF. Knockout ablates prefixes $K = 0, \ldots, 15$ of the frozen ranking. Zero ablation replaces the selected activations by zero; mean ablation replaces them by the corresponding discovery-set mean. Steering adds the natural-minus-literal activation difference at the Top-15 heads during generation. Main steering results compare $K = 15, \alpha = 1$ with the same instrumented backend at $K = 0, \alpha = 0$.

LLM judges. The registered evaluator panel contains Qwen3.8-Max (`qwen3.8-max`), GPT-5.5-0424 (`gpt-5.5-0424-global-idea`), Gemini-3.5 Flash (`gemini-3.5-flash-idea`), and Claude Opus 4.6 (`bedrock/claude-opus-4-6-idea`). Naturalness and adequacy use separate, position-randomized pairwise tasks: the former hides the source, whereas the latter includes it. The complete multilingual matrices and headline rates use Gemini-3.5 Flash; outputs from different judges remain judge-specific rather than being averaged.

For every sentence and knockout condition, five independent seeds each produce 20 generated completions. Within each seed, we estimate a five-member generated class using 100 uniformly sampled size-five subsets; sampling is without replacement within each draw. Each subset is scored against the fixed five-member counterfactual class, after which results are aggregated across draws and seeds. This equal-cardinality bootstrap uses the 20 generations while avoiding the $\log(20/5)$ raw log-sum-exp advantage and reducing dependence on a single generated subset.

Evaluation and uncertainty. Naturalness judgments hide the source and compare only the two target-language outputs. Adequacy judgments additionally show the source, but no reference. Win rates exclude ties and are therefore reported together with win/loss/tie counts. MetricX deltas are baseline minus steered because lower is better; COMET deltas are steered minus baseline because higher is better. Reported intervals use paired sentence-cluster bootstrap resampling, preserving repeated prompts and seeds from the same source unit.

Numerics and provenance. The formal Qwen3.5-4B localization and steering-direction extraction use float32 activations. Where generation kernels require bfloat16, clean-model teacher-forced scoring returns to float32.

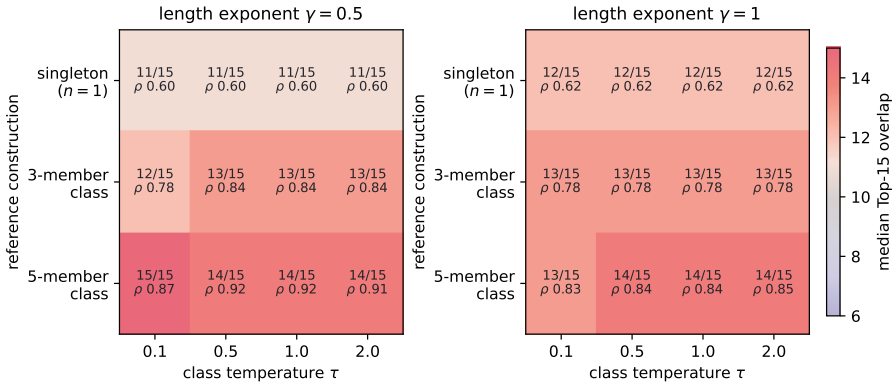

Figure 7: **Equi-DIFF sensitivity under length normalization.** Rows vary the completion-set size; $n = 1$ summarizes repeated singleton natural/literal draws. Cells report median Top-15 overlap and full-ranking Spearman correlation with the registered $(\gamma, \tau, n) = (0, 1, 5)$ ranking. These panels isolate how length normalization changes spatial localization at each completion-set size.

C Frozen Readout Validation

We validate the readout through two controlled comparisons: **(1)** word-level translation-head localization versus complete-sequence naturalness localization, and **(2)** singleton versus equivalence-class references within the complete-sequence naturalness readout. The former changes the behavioral target; the latter holds the target and scoring rule fixed while varying completion-set size. Both comparisons are frozen before the multilingual intervention study. The main text reports the frozen Top-15 comparisons. Figures 4 and 7 provide the repeated completion-choice sensitivity analysis.

Cross-ranking causal validation. We cross the frozen word-translation and five-completion naturalness Top-15 rankings with both behavioral endpoints at $K \in \{5, 10, 15\}$. Ten controls per ranking preserve layer and attention type at every rank, exclude the union of the two experimental Top-15 sets, and are frozen before intervention. Because the word logit margin and sequence Equi-DIFF have different units, each endpoint effect is divided by its bootstrap sample's unablated between-unit standard deviation before forming the interaction. Figure 8 reports the complete preregistered result.

The token endpoint uses the same frozen 132-word corpus from which the legacy word ranking was derived, whereas the 15-sentence naturalness endpoint is held out. The test establishes endpoint-selective efficacy on the frozen word endpoint and held-out generalization on the naturalness endpoint. Intervention timing follows the behavior: the word endpoint edits the final prompt position, while sequence generation edits prefill and decode positions. Both rankings are crossed within each endpoint, and zero ablation is the primary comparison to avoid an additional endpoint-specific mean-bank interpretation.

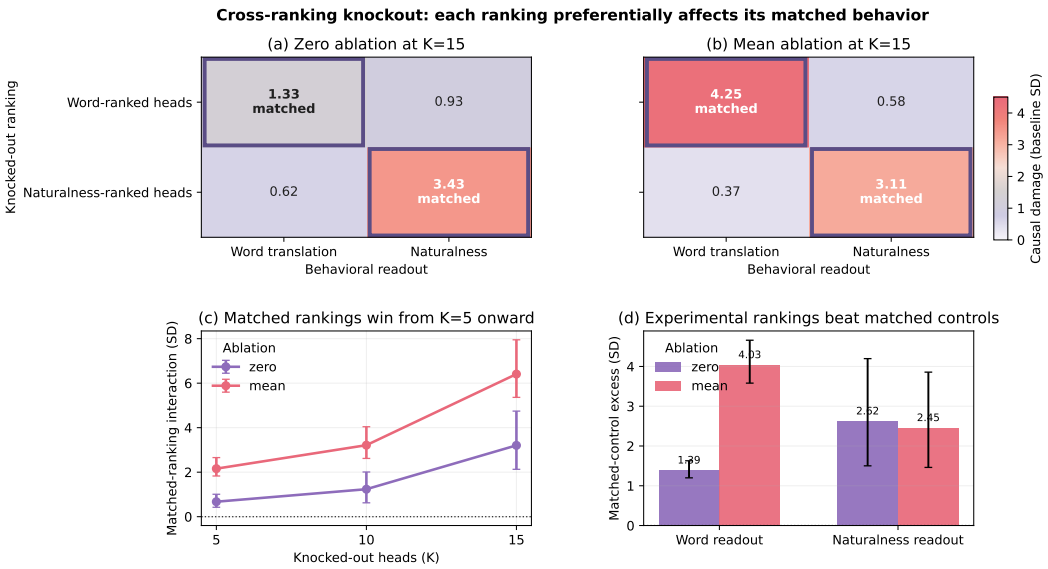

Figure 8: **Complete cross-ranking causal analysis.** (a–b) $K = 15$ readout-by-ranking crossover under zero and mean ablation. (c) Interaction across $K \in \{5, 10, 15\}$. (d) Experimental rankings versus ten layer/type-matched controls. Error bars are paired-bootstrap 95% intervals.

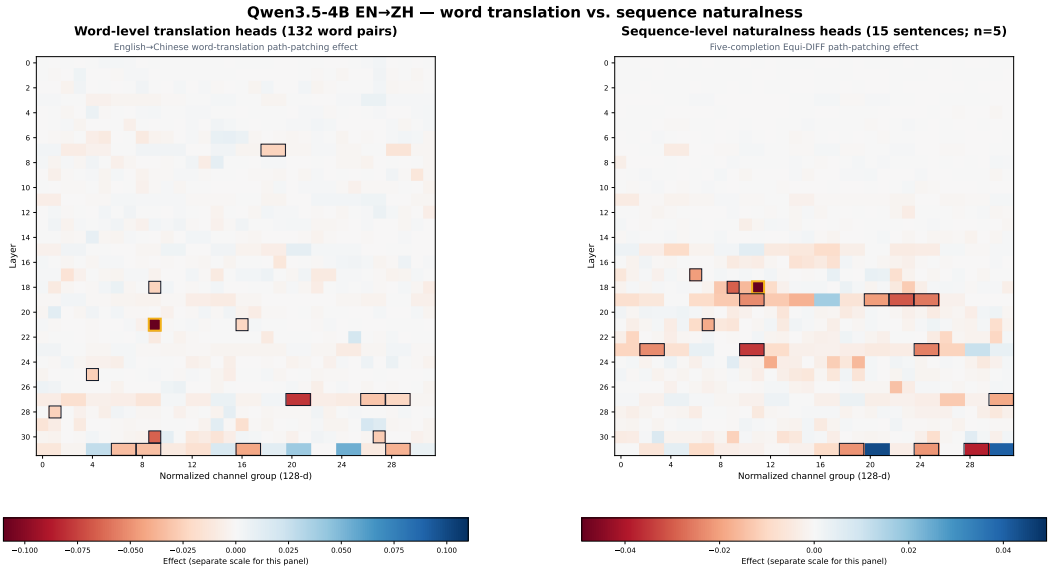

Figure 9: **Word-level translation heads versus sequence-level naturalness heads.** The left panel localizes English–Chinese word translation over 132 word pairs; the right localizes naturalness over 15 sentences using five-member completion classes. Signed AP effects use separate scales, so the comparison concerns head rankings rather than raw magnitudes. Only $2/15$ Top-15 heads overlap.

All-direction localization. Figure 10 reports the complete formal AP heatmaps of patching-induced ΔEqui-DIFF for all three models. All panels use a shared symmetric scale and outline the frozen Top-15 units. Under the paper's ranking convention, negative (red) units support the natural-style contrast and positive (blue) units the reverse. Qwen full-attention heads span two normalized 128-d cells. The upstream artifacts use the historical label ΔEC-LLR for the cardinality-normalized score contrast called Equi-DIFF in this paper.

Figure 11 gives the complementary signed macro means. For each model and attention-head coordinate, the twelve frozen direction-level effects receive equal weight. Both panels use the same symmetric $[-0.09, +0.09]$ scale as the main-text Gemma aggregate, so color depth is directly comparable across the three model-level summaries. Because opposite-signed effects cancel, these panels emphasize shared structure rather than direction-specific heads.

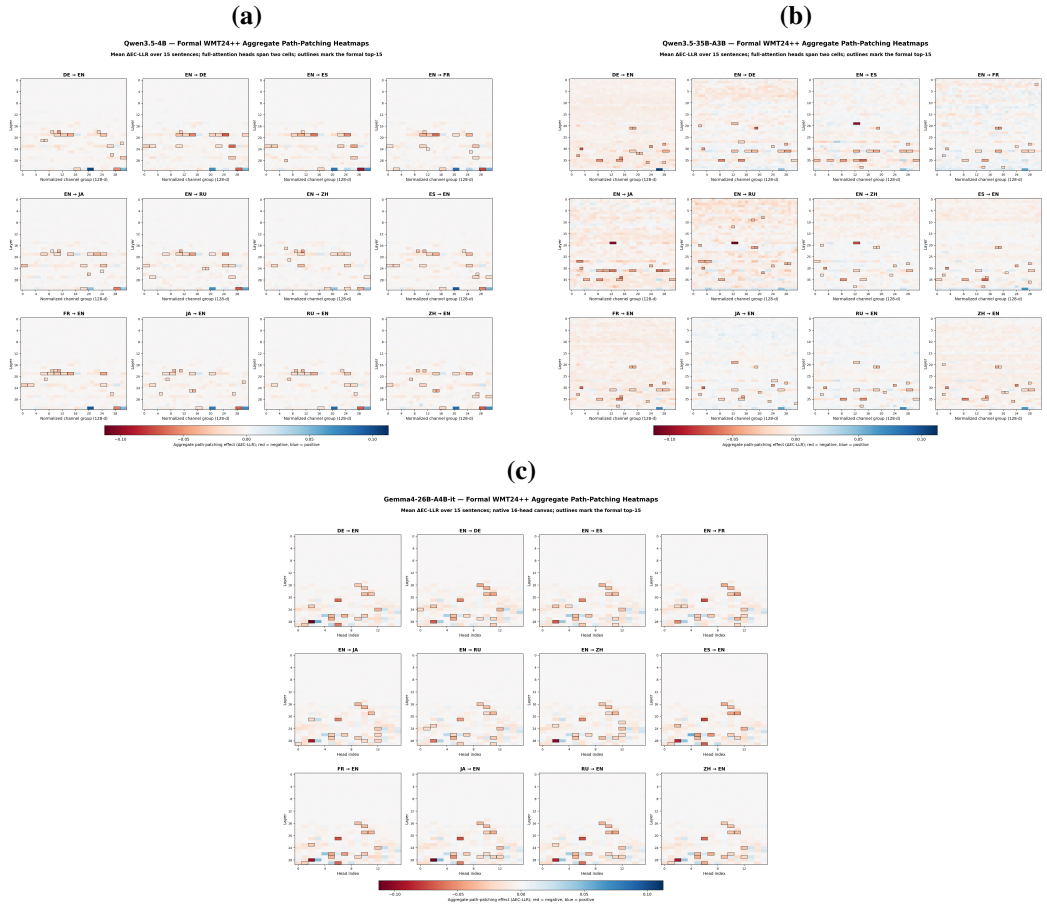

Figure 10: **Twelve-direction Equi-DIFF localization.** (a) Qwen3.5-4B, (b) Qwen3.5-35B-A3B, and (c) Gemma-4-26B-A4B-IT. Each cell shows the patching-induced change in generated-class Equi-DIFF; all panels use a shared symmetric scale, and outlines mark the frozen Top-15 units.

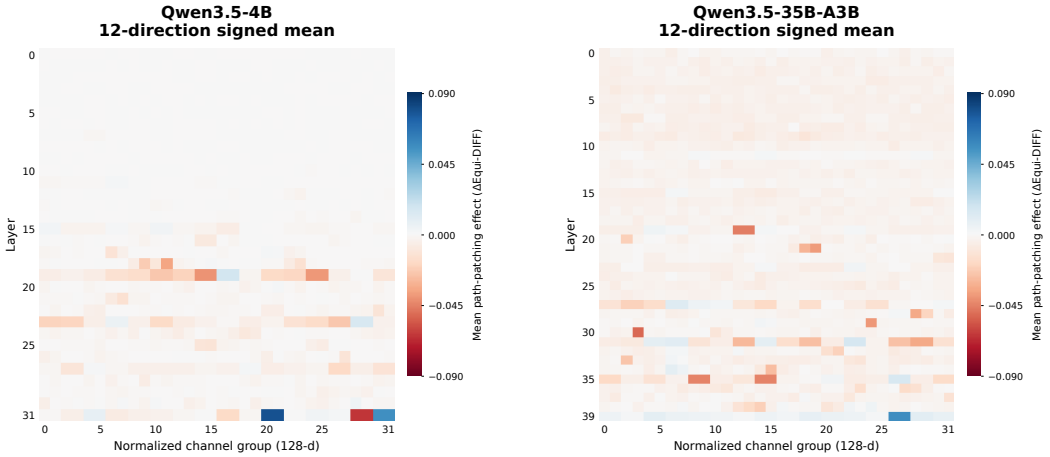

Figure 11: **Qwen twelve-direction aggregate localization.** Qwen3.5-4B (left) and Qwen3.5-35B-A3B (right) show equal-weight signed means across all twelve directions. Red is negative and blue is positive; both use the shared $[-0.09, +0.09]$ range.

Proportional head reuse by translation orientation. Finding 2 uses Top-1% as its primary cross-model budget. The English-source and English-target groups each contain 15 unordered within-group pairs among six directions. The cross-group column contains all 36 pairs between the two orientations, while matched reverse contains the six English-to-language versus language-to-English pairs for the same non-English language.

Model	Direction-pair group	Overlap@K	Jaccard	Circuit cosine	Full-rank ρ
Qwen3.5-4B	English-source	0.548	0.383	0.486	0.357
	English-target	0.615	0.450	0.638	0.554
	Cross-orientation	0.525	0.364	0.477	0.420
Qwen3.5-35B-A3B	English-source	0.522	0.365	0.682	0.213
	English-target	0.722	0.576	0.555	0.238
	Cross-orientation	0.546	0.387	0.438	0.188
Gemma-4-26B-A4B-IT	English-source	0.880	0.800	0.911	0.758
	English-target	0.813	0.695	0.824	0.775
	Cross-orientation	0.517	0.356	0.600	0.749

Table 4: **Top-1% agreement by direction-pair group.** Overlap@K and Jaccard compare selected head identities; circuit cosine compares the summed residual-write vectors of those Top-1% heads. Spearman ρ compares the complete head rankings and is therefore not restricted to the Top-1% prefix. English-source and English-target each average 15 pairs; cross-orientation averages 36 pairs.

D MULTILINGUAL KNOCKOUT

Figure 12 adds panels (b), (c), and (e) to the compact main-text selection; faint curves identify the twelve translation directions, while thick curves are equal-weight macro averages. For Qwen3.5-4B, dashed direction-level and charcoal macro curves show random-all rankings sampled from the formal Top-15 complement. These are the frozen formal curves: all pipeline stages passed, their run IDs match the matrix configuration, and the exact score-file hashes match the paper manifest.

At $K = 0$, the plotted quantity compares a five-sample bootstrap from newly generated unablated outputs with the fixed five-member counterfactual class under the reference prompt. The reference construction instead compares the fixed reference and counterfactual classes. Its sign therefore measures generated-class calibration, while causal effects use paired within-direction changes.

The curves use each direction's frozen ranking. The Qwen overview pairs the signed twelve-direction localization mean with raw and baseline-relative Equi-DIFF curves and official-reference MetricX curves for both ablation modes.

After ablating the model-specific top 1% of heads, mean Equi-DIFF loss for English-to-non-English versus non-English-to-English is 3.414 versus 1.585 for Qwen3.5-4B and 2.989 versus 1.515 for Qwen3.5-35B-A3B. Gemma shows the opposite orientation: non-English-to-English loses 20.785 units versus 8.470 for English-to-non-English. A paired bootstrap that resamples the six language pairs confirms the interaction. The English-to-non-English minus non-English-to-English contrast has 95% intervals of $[0.466, 3.447]$ and $[0.613, 3.640]$ for Qwen3.5-4B mean and zero ablation, and $[0.166, 3.054]$ and $[0.084, 2.799]$ for Qwen3.5-35B-A3B. Gemma's corresponding intervals are $[-24.680, -4.108]$ and $[-17.432, -4.575]$. Every interval remains entirely on one side of zero, meaning that the orientation difference persists under resampling rather than being driven by one language pair.

For the same top 1% of heads, four-metric mean relative quality loss is 3.82% versus 0.57% for Qwen3.5-4B, -1.17% versus -1.12% for Qwen3.5-35B-A3B, and 1.92% versus 9.96% for Gemma, with English-to-non-English listed first. Negative loss denotes a slight average quality improvement. These smaller quality changes show that the directional Equi-DIFF effect is already visible after this sparse intervention.

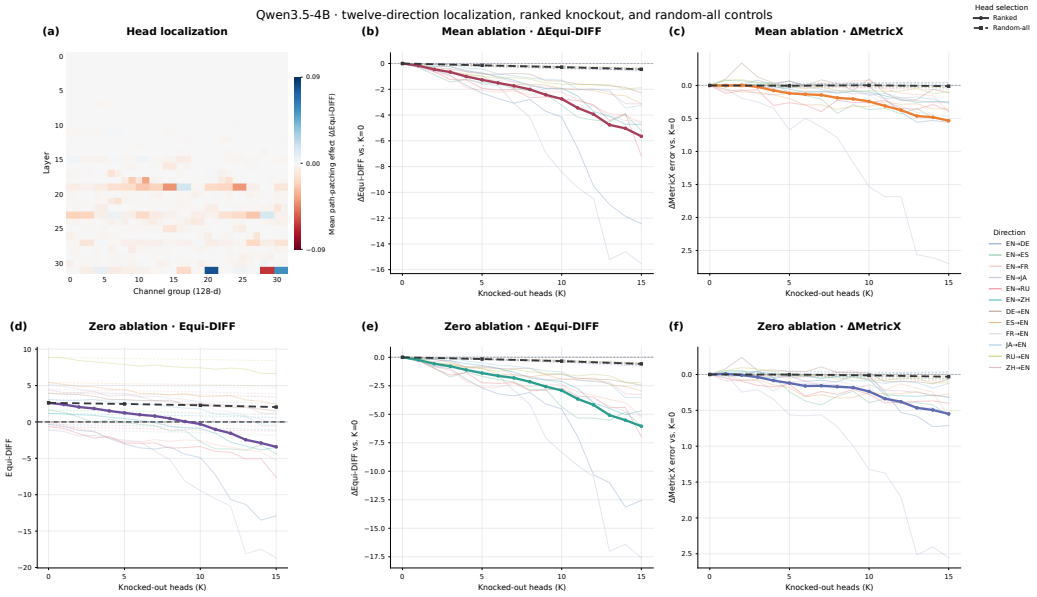

Figure 12: **Complete Qwen3.5-4B localization and twelve-direction held-out knockout.** (a) Equal-weight signed mean path-patching effect. (b–c) Baseline-relative Equi-DIFF and MetricX error under ranked mean ablation. (d) Raw Equi-DIFF under ranked zero ablation. (e–f) Baseline-relative Equi-DIFF and MetricX error under ranked zero ablation. Dashed curves in all knockout panels are random-all controls; solid curves are ranked. Random-all MetricX uses the shared formal $K = 0$ baseline. Both MetricX axes increase downward, so quality degradation is plotted as a falling trajectory.

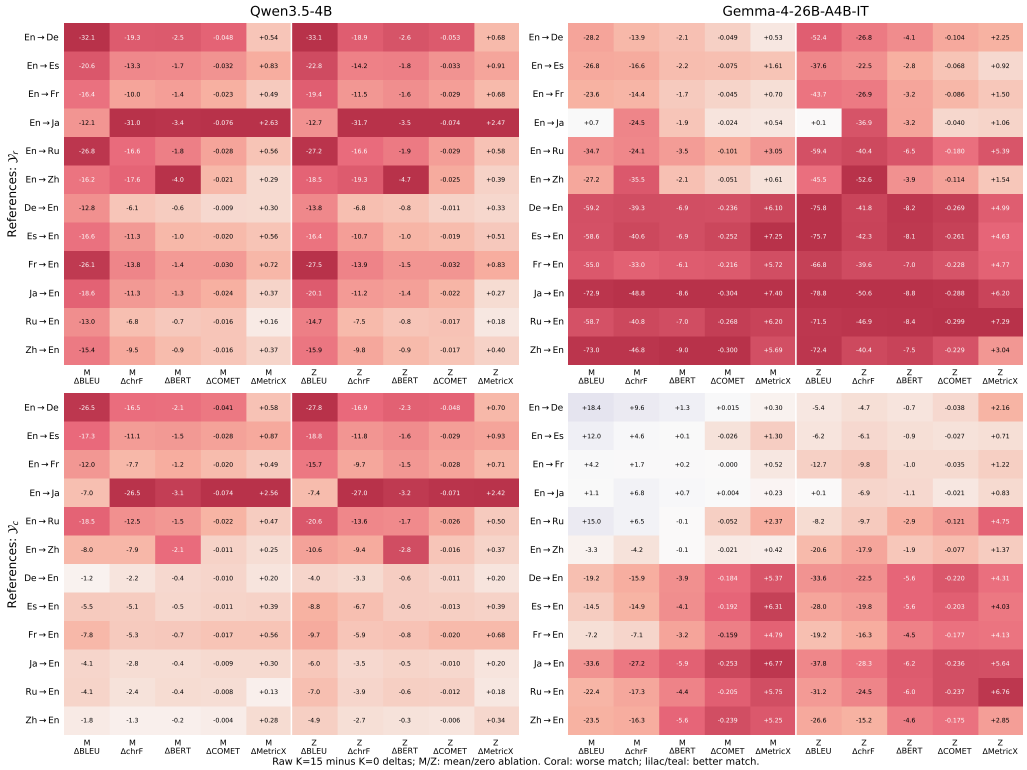

Figure 13: **Qwen3.5-4B and Gemma reference-class changes after Top-15 knockout.** The upper and lower rows score the same generations against $\mathcal{Y}_r$ and $\mathcal{Y}_c$. Cells report raw $K = 15 - K = 0$ deltas; colors are normalized within each model–metric block. M/Z denote mean/zero ablation. Coral denotes a worse class match and lilac/teal a better match.

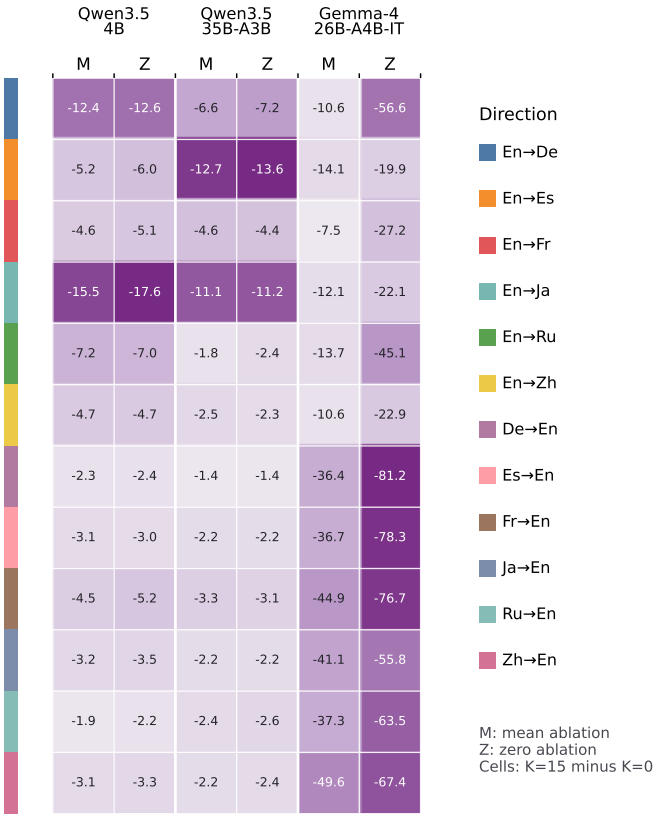

Figure 14: **Three-model Top-15 Equi-DIFF endpoint changes.** Each cell reports the raw $K = 15 - K = 0$ generated-class Equi-DIFF change for mean (M) or zero (Z) ablation. Rows are the twelve translation directions, identified by the left color strip and right-hand legend. Colors are normalized separately within each model because absolute sequence-score calibration differs across models; purple denotes a decrease in natural-over-literal preference. All 72 model–direction–mode cells are negative.

E Steering Examples and Evaluation Notes

E.1 English–Chinese evaluation notes

Main-text Table 3 jointly reports the complete three-model decisive-win-rate matrix and paired automatic-metric changes. The underlying W/L/T counts aggregate three prompt variants and remain in the frozen evaluation artifact.

Illustrative steering cases. Tables 5 and 6 show examples selected from the held-out WMT24++ test set to make the steering differences easy to inspect. They compare the instrumented $K = 0, \alpha = 0$ baseline with Top-15 steering at $\alpha = 1$; paired naturalness and adequacy changes on the 1–7 scale are reported beside each example.

Model / direction	Field	Text
Qwen3.5-4B, De→En ΔNat. $= +5$, ΔAdeq. $= 0$	Source	So sieht grad mein Badezimmer im Obergeschoss aus.
	$K = 0$	So looks my bathroom on the upper floor right now.
	Steered	That's how my bathroom on the upper floor looks right now.
Qwen3.5-35B-A3B, En→Zh ΔNat. $= +5$, ΔAdeq. $= +3$	Source	"Well, I'll stand defiantly for the Federation," Orson said defiantly. "I'll not let those terrorist scare me."
	$K = 0$	"好吧，我将为联邦defiantly 抗争，"奥森defiantly 说道，"我绝不会让那些恐怖分子吓倒。"
	Steered	"好吧，我会为联邦挺身而出，无所畏惧。"奥森语气坚定地表示，"那些恐怖分子休想吓倒我。"

Table 5: **Neutral-prompt steering examples.** Instrumented baselines and Top-15, $\alpha = 1$ outputs are shown with paired naturalness and adequacy changes.

Model / prompt / direction	Field	Text
Qwen3.5-4B, natural, De→En ΔNat. $= +4$, ΔAdeq. $= 0$	Source	@user18 Womit wird dein Gesicht verglichen?
	$K = 0$	With what are you comparing your face?
	Steered	What are you comparing your face to?
Qwen3.5-4B, literal, De→En ΔNat. $= +5$, ΔAdeq. $= +5$	Source	"So schlimm war es doch gar nicht, oder, Kayel?"
	$K = 0$	So terrible it was, or not, Kayel?
	Steered	"It wasn't that bad, was it, Kayel?"
Qwen3.5-35B-A3B, natural, En→Zh ΔNat. $= +4$, ΔAdeq. $= +4$	Source	"Well, I'll stand defiantly for the Federation," Orson said defiantly. "I'll not let those terrorist scare me."
	$K = 0$	"好吧，我将为联邦defiantly 抗争，"奥森defiantly 说道，"那些恐怖分子休想吓倒我。"
	Steered	"好吧，我会为联邦挺身而出，无所畏惧。"奥森语气坚定地表示，"那些恐怖分子吓不倒我。"
Qwen3.5-35B-A3B, literal, En→Zh ΔNat. $= +6$, ΔAdeq. $= +3$	Source	Permission to relax.
	$K = 0$	放松许可。
	Steered	放松一下，看个电视。
Gemma-4-26B-A4B-IT, literal, Ja→En ΔNat. $= +6$, ΔAdeq. $= +4$	Source	ミニチュアもっと作ってみた
	$K = 0$	Miniature more tried making.
	Steered	I tried making even more miniatures.

Table 6: **Natural- and literal-prompt steering examples.** Held-out outputs compare the instrumented baseline with Top-15 steering at $\alpha = 1$ across all three models.