Same Question, Shared Attention Groups? A Causal Study of Cross-Lingual Generalization

Anonymous EMNLP submission

Abstract

Multilingual language models can answer certain questions correctly in many languages, but we still do not know whether they rely on shared mechanisms or separate language-specific routes to answer these parallel questions. This gap limits our understanding of how multilingual generalization works within these models. We address this gap by using causal interventions to compare which components support correct answers for parallel question-answering examples across languages. Across two 8B models, two multilingual benchmarks, and five languages, we find that matched questions show more similar component-importance patterns than mismatched questions within and across languages. Language identity explains little of the variation in component importance; most variation is tied to the question itself or to question-language interactions. We make this reuse concrete in a two-hop reasoning task, showing that causal intervention can transfer an inferred intermediate concept across languages. Together, these results suggest that correct multilingual behavior often depends on shared, question-specific mechanisms, while still allowing language-specific computation.

1 Introduction

Multilingual language models are used as general-purpose systems by users across many languages. They are intended to support cross-lingual transfer and expand access to language technologies for diverse communities. However, their behavior remains uneven across languages: performance can differ from one language to another, and knowledge may be encoded or retrieved inconsistently across languages. The internal mechanisms underlying these behaviors remain poorly understood. Even when a model answers the same question correctly across languages, this shared outcome does not reveal how the model arrives at the answer internally.

Prior work on multilingual internal organization has largely relied on representational analysis, showing that multilingual models can exhibit partially shared internal states and may pass through a dominant or latent language for non-English inputs (Lim et al., 2025; Wendler et al., 2024; Zhong et al., 2025). A separate line of work shows that attention heads vary in importance and that many can be pruned while preserving aggregate task performance (Michel et al., 2019; Voita et al., 2019; Ma et al., 2021). Neither line answers a finer-grained causal question: when the same question is asked in different languages, do the same attention groups support the answer, or do different languages rely on different groups?

We study this question in cases where the model's external behavior is already aligned across languages: parallel question-answering examples that the model answers correctly in every language considered. This setting lets us ask whether behavioral agreement is supported by similar attention-group usage patterns, or whether different languages arrive at the same answer through different routes.

We approach this question by causally intervening on attention groups and comparing their importance across parallel inputs through activation patching. Across two 8B models, two multilingual question-answering benchmarks, and five languages, we find that matched questions produce more similar attention-group usage signatures than mismatched questions within or across languages. Language identity explains little of the variation in attention-group importance; most variation is tied to the question itself or to question-language interactions. These results suggest that correct cross-lingual answers often rely on shared question-dependent attention structure, while still allowing language-specific processing.

We further analyze a two-hop reasoning task to ask whether attention-group reuse across lan-

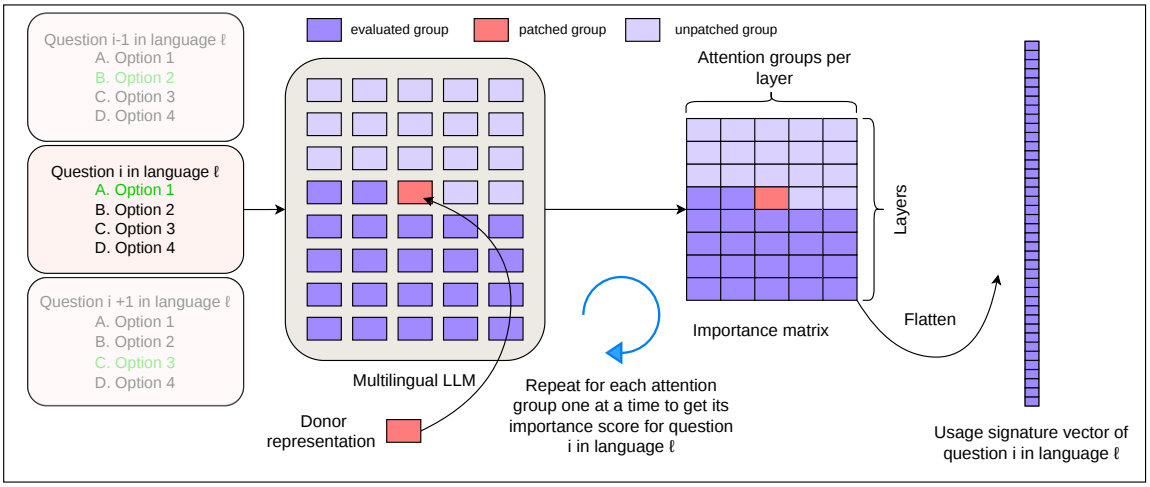

Figure 1: Overview of the attention-group usage signature construction. For each question in language ℓ, we patch one attention group at a time with a language-matched donor representation and measure the resulting change in the model's preference for the correct answer. Repeating this over layers and attention groups yields an importance matrix, which we flatten into a usage signature vector for that question-language pair.

guages extends beyond final-answer agreement to an internal reasoning step. In this task, the model must use an implicit intermediate concept to map an input to the final answer, for example using source $\rightarrow$ intermediate concept $\rightarrow$ target, without that concept being stated explicitly. We find that the same localized attention groups recur across languages, and that patching these groups across languages shifts the model's output in the direction predicted by the donor prompt while preserving the recipient prompt's language and answer type. This provides a targeted example of cross-lingual attention-group reuse supporting an internal reasoning step, not just a shared final answer.

Our contributions are as follows:

- We formulate multilingual alignment as *causal reuse*: whether semantically matched inputs across languages are supported by overlapping causally important components, rather than only similar outputs or activations.
- We provide evidence that attention-group importance is partly organized around the problem being solved, not simply around the input language. This helps explain how models can share mechanisms across languages while still exhibiting language-specific behavior.
- We show, through a two-hop reasoning case study, that cross-lingual attention-group reuse can support an intermediate reasoning step, not only final-answer agreement.

Overall, our results provide a step toward understanding the internal mechanisms that support cross-lingual generalization. They also suggest a path for future multilingual safety and editing interventions: first identify components reused across semantically matched inputs, then test whether interventions on those components transfer across languages without unwanted effects on unrelated inputs.

2 Related Work

Multilingual internal organization. Prior work has studied multilingual models largely through representational and probing analyses. These studies show that multilingual models contain both shared and language-specific structure, with language-specific processing sometimes limiting transfer and consistency (Lim et al., 2025). Other work finds that multilingual transformers may pass through dominant or latent internal languages, often with English-leaning intermediate states for non-English inputs (Wendler et al., 2024; Zhong et al., 2025). Mechanistic analyses of cross-lingual factual inconsistency further suggest that knowledge can be represented in relatively language-independent concept spaces before shifting to more language-specific spaces near the output (Wang et al., 2025). Related work also finds that models behave more similarly internally across languages on knowledge-free reasoning tasks than on tasks requiring knowledge retrieval (Hu et al.,

2025). Together, these findings show that multilingual computation involves both shared and language-dependent structure. Our work asks a complementary causal question: whether semantically matched inputs across languages depend on overlapping model components.

Attention components and multilingual specialization. Attention heads are a common unit of component-level analysis because they can be ablated, pruned, or otherwise intervened on. Prior work shows that many heads can be removed with limited effect on aggregate performance, while smaller subsets often carry disproportionate functional importance (Michel et al., 2019; Voita et al., 2019). In multilingual settings, head pruning, head-importance, and subnetwork analyses have been used to study cross-lingual transfer, multilingual task performance, language-specific modularity, and the distinction between language-specific and language-general components (Ma et al., 2021; Choenni et al., 2024; Liu et al., 2026). These studies show that attention components can be redundant, specialized, or language-sensitive. However, language-level importance scores, sparse subnetworks, and task-level pruning do not show whether the same components support specific semantically matched instances in different languages. We instead compare component importance at the level of matched question instances.

Causal interventions and attention-group units. Mechanistic interpretability uses interventions such as activation patching, causal tracing, causal mediation analysis, path patching, attribution patching, and distributed alignment search to test whether internal states or components affect model behavior (Meng et al., 2022; Goldowsky-Dill et al., 2023; Syed et al., 2024; Geiger et al., 2024). These methods have been used to identify behavior-specific circuits (Wang et al., 2023; Conmy et al., 2023) and are increasingly being considered in multilingual settings (Long, 2025). The models studied here use grouped-query attention, where multiple query heads share key-value projections (Ainslie et al., 2023). This makes attention groups a natural intervention unit: each group corresponds to a shared key-value pathway that can be patched consistently across prompts. Our usage signatures therefore measure causal sensitivity over layers and attention groups, leaving open the role of MLPs, residual-stream features, and cross-layer circuits.

Latent intermediate computations. A related line of work studies whether language models use identifiable intermediate variables during in-context learning. Latent Concept Disentanglement analyzes a monolingual two-hop setting in which the intermediate concept is unstated, and localizes attention components involved in mediating that concept (Hong et al., 2026). We adapt this setup to a multilingual two-hop setting as a concrete case study of attention-group reuse across languages. Unlike the usage-signature analysis, which measures aggregate overlap across many benchmark questions, this experiment asks whether localized groups can transfer an inferred intermediate concept across languages while preserving the recipient prompt's language and answer type.

3 Methodology

We measure cross-lingual reuse of attention groups in four steps. First, we define a per-question attention-group usage signature using activation patching (§3.1). Second, we compare signatures for matched questions across languages against shuffled controls (§3.2). Third, we decompose importance variation into language, question, and language-question interaction components (§3.3). Finally, we test whether localized attention groups can transfer an inferred intermediate concept across languages in a two-hop task (§3.4).

3.1 Attention-group usage signatures

We intervene at the attention-group level (Ainslie et al., 2023). In a transformer layer, the attention module first computes attention outputs from query, key, and value projections. These outputs are concatenated across heads, passed through the attention output projection, and then added back to the residual stream. In grouped-query attention, multiple query heads share one key-value head, so the pre-output-projection attention result can be partitioned into groups corresponding to shared key-value pathways. We define an attention group as one such pre-output-projection slice. Patching a group replaces this slice at the final token position with a donor activation, leaving the other attention groups, the output projection, and the rest of the layer unchanged. Figure 1 illustrates the full procedure for constructing an attention-group usage signature. Appendix B gives the intervention site, signature dimensionality, donor construction, and patching procedure.

Let $a^{\ell}_{i,l,g}$ be the output of group g in layer l at the final token position for question i in language ℓ, captured before the attention output projection. Let $\mathcal{D}_\ell$ be the retained multilingual-correct questions in language ℓ. For each dataset, model, and language, we form a language-matched donor by averaging over $\mathcal{D}_\ell$:

$$d_{\ell,l,g} = \frac{1}{|\mathcal{D}_\ell|} \sum_{i \in \mathcal{D}_\ell} a^{\ell}_{i,l,g}. \tag{1}$$

This donor is language-matched but not question-matched: it neutralizes question-specific information at the patched site while preserving broad activation statistics for the recipient language. We use this mean-replacement intervention instead of zeroing the group output because it better preserves the scale and distribution of activations in the recipient language.

For each question-language pair, we rerun the model once per candidate group and replace that group's final-token output with $d_{\ell,l,g}$. We then measure how much the patch reduces the model's preference for the correct answer. For option label c, let $z^{\text{base}}_{i,c}$ and $z^{\text{patch}}_{i,l,g,c}$ be the baseline and patched label log-probability scores, let y_i be the correct label, and let C be the number of options. We define the correct-vs-average-incorrect margin as

$$m(z, y_i) = z_{y_i} - \tfrac{1}{C-1} \textstyle\sum_{c \neq y_i} z_c, \tag{2}$$

and the group importance as

$$I_{i,l,g} = 100 \cdot \frac{m(z^{\text{base}}_i, y_i) - m(z^{\text{patch}}_{i,l,g}, y_i)}{\left|m(z^{\text{base}}_i, y_i)\right| + \varepsilon}, \tag{3}$$

with $\varepsilon = 10^{-8}$. Larger values indicate that perturbing the group more strongly weakens the model's preference for the correct answer.

To reduce answer-position confounds, we patch each question under four answer-order permutations that preserve answer contents and correct-option identity (Appendix A). We average importance over permutations and flatten over layers and groups to obtain a usage signature vector $v^{\ell}_i \in \mathbb{R}^H$, where H is the total number of attention groups.

3.2 Cross-language similarity and controls

We use usage signatures to test whether matched questions rely on similar attention groups across languages. We use *aligned* to denote comparisons between parallel versions of the same question across languages. For language pair (ℓ, ℓ') and multilingual-correct question set $\mathcal{Q}$, we define

$$S^{\text{aligned}}_{\ell,\ell'} = \tfrac{1}{|\mathcal{Q}|} \textstyle\sum_{i \in \mathcal{Q}} s\left(v^{\ell}_i,\, v^{\ell'}_i\right), \tag{4}$$

where s is Spearman correlation or Jaccard@10.

We compare this aligned similarity against two shuffled nulls. The *cross-language* null permutes question identifiers in ℓ' while preserving the language pair, model, and dataset. This breaks semantic matching across languages. The *within-language* null compares different questions within the same language. These controls test whether aligned similarity reflects the matched question itself rather than generic language, task-format, or broadly important-group effects.

We draw $B = 500$ permutations for each null and report effect sizes as the gap to the cross-language null mean:

$$\Delta_{\ell,\ell'} = S^{\text{aligned}}_{\ell,\ell'} - \tfrac{1}{B} \textstyle\sum_{b=1}^{B} S^{\text{null},b}_{\ell,\ell'}. \tag{5}$$

We compute one-sided permutation p-values as

$$p = \frac{1 + \sum_{b=1}^{B} \mathbf{1}\left[S^{\text{null},b}_{\ell,\ell'} \geq S^{\text{aligned}}_{\ell,\ell'}\right]}{B+1}. \tag{6}$$

We treat aggregate aligned-null gaps as the main effect-size summary. We also compute Spearman similarity on single-layer slices to locate where reuse appears in the network; we report Jaccard@10 only at the whole-signature level.

3.3 Variance decomposition of importance

Signature similarity shows whether aligned pairs are closer than shuffled pairs, but it does not reveal what organizes the underlying relative attention-group importance rankings. To separate language-driven and question-driven structure, we therefore decompose the rank-normalized importance tensor directly.

We arrange importance values into a balanced tensor $I_{\ell,q,h}$ indexed by language ℓ, question q, and flattened attention group h. Because raw patching effects can differ in scale across language-question pairs, we convert scores within each pair (ℓ, q) into rank percentiles over groups:

$$X_{\ell,q,h} = \frac{\text{rank}_{\ell,q}(I_{\ell,q,h}) - 1}{H - 1} \in [0, 1]. \tag{7}$$

Thus the decomposition uses relative group importance rather than raw intervention magnitude.

For each group h, let $Y_{\ell,q} = X_{\ell,q,h}$ be the $L \times Q$ table over languages and questions. Let μ be the grand mean, μ_ℓ the mean over questions for language ℓ, and μ_q the mean over languages for question q. We define

$$a_\ell = \mu_\ell - \mu, \tag{8}$$

$$b_q = \mu_q - \mu, \tag{9}$$

$$r_{\ell,q} = Y_{\ell,q} - \mu - a_\ell - b_q, \tag{10}$$

corresponding to language, question, and interaction components. On the balanced grid, the corresponding sums of squares are

$$SS_{\text{lang},h} = Q \sum_\ell a_\ell^2, \tag{11}$$

$$SS_{\text{q},h} = L \sum_q b_q^2, \tag{12}$$

$$SS_{\text{int},h} = \sum_{\ell,q} r_{\ell,q}^2. \tag{13}$$

We aggregate across groups before forming fractions:

$$F_c = \frac{\sum_h SS_{c,h}}{\sum_h SS_{\text{tot},h}}, \quad c \in \{\text{lang}, q, \text{int}\}. \tag{14}$$

The fractions F_{lang}, F_{q}, and F_{int} quantify how much relative importance variation is explained by language identity, question identity, and language-question interaction. This tells us whether attention importance is mainly organized by language, by the problem being solved, or by particular question-language pairs.

3.4 Multilingual latent-country case study

The usage-signature analysis measures aggregate reuse across benchmark questions. We complement it with a targeted test of whether localized groups can transfer an inferred intermediate concept across languages in a controlled two-hop task. We instantiate this general setup as a latent-country task adapted from the latent two-hop setup of Hong et al. (2026): each prompt follows source $\rightarrow$ country $\rightarrow$ target, where the country is the hidden bridge variable. Normal and alternative prompts use different source and target types, so successful transfer cannot be explained by copying the donor answer string.

For a normal prompt p^{ℓ}_{norm} and an alternative prompt $p^{\ell'}_{\text{alt}}$, we write $\ell' \rightarrow \ell$ for patching from language ℓ' into language ℓ. The evaluation target is the *type-corrected alternative answer*: the answer associated with the alternative prompt's latent country, expressed in the target type and language of the normal prompt. Figure 2 illustrates this target definition.

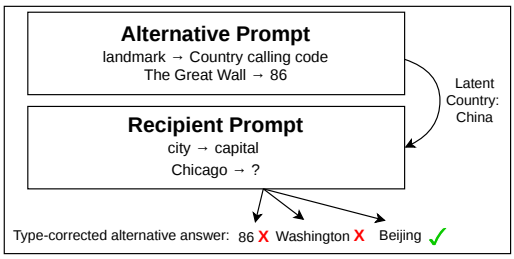

Figure 2: Illustration of the type-corrected alternative answer. The alternative prompt identifies the latent country, China, through a landmark-to-calling-code relation, while the recipient prompt specifies the target answer type, capital. The type-corrected alternative answer is therefore Beijing, not the donor answer string 86 or the recipient prompt's original answer Washington.

Localization. For each candidate attention group m, we cache its final-token output on the alternative prompt and substitute it into the same site in the normal prompt:

$$a_m(p^{\ell}_{\text{norm}}) \leftarrow a_m(p^{\ell'}_{\text{alt}}). \tag{15}$$

If group m carries information about the hidden country, this patch should move the model from the normal answer toward the type-corrected alternative answer. We perform localization separately for each language pair to test whether the same high-scoring groups recur across pairs.

Localization score. Let Δ_i^{base} be the unpatched logit margin between the normal answer and the type-corrected alternative answer for pair i, and let $\Delta_i^{\text{patch},m}$ be the same margin after patching group m. We score each group by

$$\text{Score}_m = \frac{1}{N} \sum_{i=1}^{N} \frac{\Delta_i^{\text{base}} - \Delta_i^{\text{patch},m}}{\Delta_i^{\text{base}}}. \tag{16}$$

Larger values mean that patching group m moves the model more strongly toward the country-consistent alternative answer.

Patching evaluation. For each condition, we patch the top-scoring group or a small recurrent set of top-scoring groups and evaluate 100 sampled normal-alternative prompt pairs. We report top-1 and top-10 vocabulary-rank rates, as well as mean

and median rank, for the first token of the type-corrected alternative answer. These metrics test whether patching promotes the country-consistent target token, rather than merely lowering confidence in the original normal answer.

4 Experimental Setup

We design the experiments to answer two questions: whether parallel questions answered correctly across languages rely on similar attention groups, and whether such reuse can support an intermediate reasoning step under cross-lingual patching.

Models. We study Qwen3-8B (Yang et al., 2025) and Llama-3.1-8B (Grattafiori et al., 2024), two open-weight decoder-only models with grouped-query attention.

Languages. We evaluate on English, Spanish, German, Russian, and Arabic, covering languages that vary in script, family, and proximity to English: Spanish and German are closer Indo-European languages, Russian adds a Cyrillic-script Slavic language, and Arabic adds a Semitic language written in Arabic script.

Benchmarks. We use two multilingual benchmarks of parallel multiple-choice questions: Belebele (Bandarkar et al., 2024), a passage-based reading-comprehension benchmark, and MMLU-ProX (Xuan et al., 2025), a multilingual reasoning benchmark. In both benchmarks, questions and answer options are parallel across languages. We elicit answers as option labels rather than free-form answer strings: Belebele uses labels A–D, and MMLU-ProX uses the first C labels from A–J for questions with C options. This keeps the output interface fixed across languages while the question and option text vary by language. Exact prompt templates and answer-extraction rules are given in Appendix A.

Multilingual-correct filter. To compare component usage only when the model solves the task in every language, all similarity analyses are restricted, per model, to questions answered correctly in all five languages. Retained sample sizes are reported in Section 5 (Table 1). This filter excludes language-specific failures, where a model answers a question correctly in some languages but not others; analyzing those cases is left to future work.

Dataset	Model	Retained	%
Belebele	Qwen3-8B	478 / 900	53.1
Belebele	Llama-3.1-8B	380 / 900	42.2
MMLU-ProX	Qwen3-8B	1839 / 11829	15.5
MMLU-ProX	Llama-3.1-8B	1420 / 11829	12.0

Table 1: Retained multilingual-correct questions per dataset and model. Retention is well above the random-guessing expectation in every condition ($E[X] = 0.88$ for Belebele, $E[X] = 0.768$ for MMLU-ProX).

Latent-country dataset. For the case study in §3.4, we construct a controlled multilingual two-hop dataset in which the model must infer an latent country before producing the final answer. Prompts are generated from aligned language files covering the same countries and relations. The normal prompt uses `city→capital`, while the alternative prompt uses a different source and target type, such as `landmark→country code` or `university→country code`. This design lets us test whether patching localized attention groups transfers the inferred latent country across languages, rather than copying an answer string or preserving a surface pattern. Construction and scoring details are given in Appendix D.

5 Results

5.1 Retained multilingual-correct sample

All similarity analyses are restricted to questions that the model answers correctly in all five languages. Table 1 reports the retained sample sizes. Retention is substantially above what would be expected under independent random guessing across five languages, which gives $E[X] = 0.88$ for Belebele and $E[X] = 0.768$ for MMLU-ProX (computed per item, since MMLU-ProX option counts vary). Retention is also higher on Belebele than on MMLU-ProX in both models, consistent with MMLU-ProX being a more challenging reasoning-oriented benchmark (Xuan et al., 2025).

5.2 Aligned questions are more similar than shuffled controls

Across all four model-dataset combinations, aligned signatures are more similar than the cross-language null (Spearman $\Delta \in [0.232, 0.427]$, Jaccard@10 $\Delta \in [0.113, 0.245]$; Table 2). Per-language-pair gaps are reported in Appendix E; all gaps are positive and all one-sided permutation tests attain the minimum attainable value with $B = 500$ permutations, $p = 1/(B + 1) \approx 0.002$.

Dataset	Model	Spearman			Jaccard@10		
		Aligned	Null	Within	Aligned	Null	Within
Belebele	Llama-3.1-8B	0.448	0.191	0.196	0.513	0.354	0.361
MMLU-ProX	Llama-3.1-8B	0.569	0.142	0.146	0.529	0.284	0.286
Belebele	Qwen3-8B	0.345	0.112	0.128	0.456	0.344	0.358
MMLU-ProX	Qwen3-8B	0.395	0.104	0.112	0.413	0.275	0.283

Table 2: Multilingual signature similarity by dataset and model, averaged over the 10 cross-language pairs. *Aligned* is the mean similarity for semantically matched questions, *Null* is the mean similarity under cross-language shuffled pairing, and *Within* is the mean similarity under within-language shuffled pairing (averaged over the five languages). Per-language-pair gaps are reported in Appendix E.

The within-language shuffled control is nearly indistinguishable from the cross-language null in every condition: Spearman differs by less than 0.01, and Jaccard@10 by less than 0.02. Thus breaking question identity has a similar effect whether the comparison is cross-lingual or within-language, suggesting that the aligned-versus-null gap is tied to matched question identity rather than to language-specific generic groups alone.

Figure 3 shows that aligned cross-lingual signatures are more similar than shuffled signatures across most relative layer depths. The gap is small but positive in early layers, grows through the middle of the network, and remains large in later layers. This pattern suggests that the aligned effect is not solely a final-layer phenomenon. At the same time, the shuffled null also increases in the upper half of the model, which points to a substantial contribution from generic task structure, shared answer-selection behavior, or broadly important attention groups. We therefore interpret the layerwise result as evidence that question-specific reuse is present beyond generic format effects, rather than as evidence that format effects are absent.

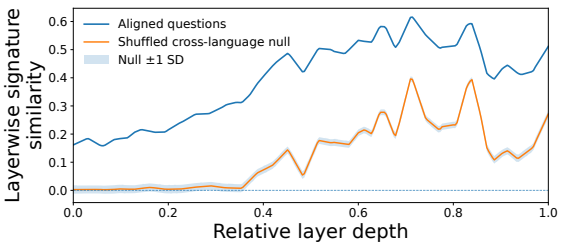

Figure 3: Layerwise cross-language signature similarity. For each relative layer depth, we compute Spearman similarity between single-layer attention-group usage signatures and average over all 10 language pairs and all four model-dataset combinations. The aligned curve compares semantically matched questions across languages. The shuffled null permutes question identities across languages while preserving the language pair, model, and dataset. The shaded band shows the null mean ± 1 standard deviation across language-pair and model-dataset combinations.

5.3 Variance decomposition: language main effects are small

The similarity results show that aligned signatures are closer than shuffled ones, but they do not say whether variation in a group's importance is organized by language identity, question identity, or particular language-question combinations. We answer this directly by decomposing the rank-normalized importance tensor as described in Section 3.3.

Table 3 reports $(F_{\text{lang}}, F_{\text{q}}, F_{\text{int}})$ for each model-dataset combination. Language main effects are small: F_{lang} is below 1.5% in every condition. The remaining variation is split between question-driven and interaction-driven components, with F_{q} accounting for roughly 41% to 60% and F_{int} for roughly 40% to 58%.

Because the decomposition uses within-pair ranks, these fractions describe variation in a group's relative importance among groups, not variation in raw intervention magnitude. The small F_{lang} therefore does not mean that language has no influence on the model's computation. It means that, after each language-question pair is normalized into a rank ordering over attention groups, language identity explains little systematic variation in those rankings.

The remaining variation is split between question identity and language-question interaction. A larger F_{q} means that parallel versions of the same question tend to assign high relative importance to similar groups across languages. A larger F_{int} means that a group's importance also depends on the particular language-question combination, which may reflect translation-specific behavior, language-contingent solution strategies, or estima-

Dataset	Model	F_{lang}	F_{q}	F_{int}
Belebele	Llama-3.1-8B	0.005	0.453	0.542
MMLU-ProX	Llama-3.1-8B	0.003	0.598	0.399
Belebele	Qwen3-8B	0.014	0.407	0.579
MMLU-ProX	Qwen3-8B	0.008	0.459	0.533

Table 3: Variance decomposition of the rank-normalized importance tensor by model and dataset, aggregated over all attention groups. F_{lang}, F_{q}, and F_{int} are the SS-weighted share of variation attributable to language identity, question identity, and language-question interaction; they sum to one. Layerwise decompositions appear in Appendix F.

Model	Prominent attention groups
Qwen3-8B	L24/G6, L31/G0
Llama-3.1-8B	L14/G6, L15/G4, L17/G5, L24/G0

Table 4: Attention groups identified by localization as carrying the latent country. For Qwen3-8B, L24/G6 appears in every condition; L31/G0 appears in every condition except monolingual Russian. For Llama-3.1-8B, the same four-group set is used in every monolingual and cross-lingual condition.

Model	Setting	T1_b	T1_p	T10_p	rank_p
Qwen3-8B	Mono	0.8	34.3	76.7	3.2
Qwen3-8B	Cross	0.5	35.6	78.0	2.4
Llama-3.1-8B	Mono	0.6	12.3	61.9	6.8
Llama-3.1-8B	Cross	0.4	12.9	64.6	5.9

Table 5: Aggregate vocabulary-rank results for the multilingual latent-country case study. T1_b and T1_p are the percentages of examples where the first token of the type-corrected alternative answer is ranked first in the full model vocabulary before and after patching. T10_p is the percentage where that token is ranked in the top 10 after patching, and rank_p is the patched median vocabulary rank. Mono aggregates 10 experiment-level conditions and Cross aggregates 16. Per-condition results are in Appendix G.

tion noise. Thus the decomposition supports a nuanced version of the similarity result: attention-group importance is not primarily organized by language main effects, but it is also not explained by question identity alone.

5.4 Multilingual latent-country case study

The case study moves from broad signature comparison to a targeted intervention test: can localized attention groups transfer a latent country signal across languages? Table 4 lists the groups identified by localization. Qwen3-8B consistently localizes L24/G6, with L31/G0 appearing in all conditions except monolingual Russian. Llama-3.1-8B consistently localizes the same four-group set across monolingual and cross-lingual conditions.

Aggregate vocabulary-rank results. Patching the prominent groups substantially promotes the type-corrected alternative answer in every model and setting (Table 5). Baseline top-1 vocabulary-rank rate is below 1% throughout, indicating that the target token is not naturally favored. After patching, top-1 rises to roughly one third of examples for Qwen3-8B and one eighth for Llama-3.1-8B, with patched top-10 reaching 62–78% and median vocabulary rank falling to single digits.

Cross-lingual patching remains effective. The monolingual and cross-lingual settings produce comparable interventions, with cross-lingual top-1 and top-10 rates at least as high as monolingual in every aggregate condition. This is consistent with a latent country signal that remains usable across donor and recipient languages under patching.

6 Conclusion and Future Work

This work frames multilingual generalization as *causal reuse*: the extent to which semantically matched inputs across languages depend on overlapping causally important components. Our results suggest that correct cross-lingual behavior is not only reflected in similar outputs or shared representations, but also in reusable attention-group structure. At the same time, reuse is not complete: substantial question-language interaction shows that shared mechanisms coexist with language-specific or translation-specific computation.

These findings have implications for multilingual intervention design. If safety, editing, or auditing methods localize and target shared components, their effects may transfer beyond the language in which they were developed. This transfer could be useful for building interventions that generalize across languages, but it also means that component-level interventions should be evaluated for unintended cross-lingual side effects.

Future work should extend this analysis beyond attention groups to MLPs, residual-stream features, and multi-layer component sets. It should also study broader language sets, open-ended generation, and cases where models succeed in one language but fail in another.

Limitations

This study focuses on attention groups as a tractable and causally interpretable unit of analysis. This choice is well matched to the grouped-query attention architecture of the models studied, but it does not cover all components that may contribute to multilingual behavior, including MLP activations, residual-stream features, and multi-layer circuits. Our main similarity and variance analyses are also restricted to questions that each model answers correctly in all five languages. This design lets us compare internal component usage when external behavior is already aligned, but it means that our conclusions apply most directly to successful multilingual generalization. Cases where a model answers correctly in some languages but not others may involve different mechanisms, weaker reuse, or language-specific failure modes.

Our experiments focus on two open-weight 8B models and five languages: English, Spanish, German, Russian, and Arabic. This setting provides variation in script, language family, and proximity to English, and the consistency of the results across two models and two benchmarks suggests that the observed patterns are not tied to a single model or dataset. At the same time, extending the analysis to additional model families, model scales, and a broader set of languages, including lower-resource and typologically diverse languages, would provide a fuller picture of how broadly these patterns hold.

Ethical Considerations

This work is intended to improve understanding and auditing of multilingual models by clarifying when correct answers across languages are supported by shared internal components. Such insights can support more reliable multilingual interpretability, debugging, and future safety or editing methods. The latent-country case study uses country-related facts only as a controlled setting for studying intermediate computation; the dataset is fact-checked and its translations are reviewed by native speakers. We do not use personal or sensitive user data, and the task is not intended to evaluate people, cultures, or countries.

References

Joshua Ainslie, James Lee-Thorp, Michiel de Jong, Yury Zemlyanskiy, Federico Lebron, and Sumit Sanghai. 2023. GQA: Training generalized multi-query transformer models from multi-head checkpoints. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4895–4901, Singapore. Association for Computational Linguistics.

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. 2024. The belebele benchmark: a parallel reading comprehension dataset in 122 language variants. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 749–775, Bangkok, Thailand. Association for Computational Linguistics.

Rochelle Choenni, Ekaterina Shutova, and Dan Garrette. 2024. Examining modularity in multilingual LMs via language-specialized subnetworks. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 287–301, Mexico City, Mexico. Association for Computational Linguistics.

Arthur Conmy, Augustine Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. 2023. Towards automated circuit discovery for mechanistic interpretability. In *Advances in Neural Information Processing Systems*, volume 36, pages 16318–16352. Curran Associates, Inc.

Atticus Geiger, Zhengxuan Wu, Christopher Potts, Thomas Icard, and Noah Goodman. 2024. Finding alignments between interpretable causal variables and distributed neural representations. In *Proceedings of the Third Conference on Causal Learning and Reasoning*, volume 236 of *Proceedings of Machine Learning Research*, pages 160–187. PMLR.

Nicholas Goldowsky-Dill, Chris MacLeod, Lucas Sato, and Aryaman Arora. 2023. Localizing model behavior with path patching.

Aaron Grattafiori et al. 2024. The llama 3 herd of models.

Guan Zhe Hong, Bhavya Vasudeva, Vatsal Sharan, Cyrus Rashtchian, Prabhakar Raghavan, and Rina Panigrahy. 2026. Latent concept disentanglement in transformer-based language models. In *The Fourteenth International Conference on Learning Representations*.

Peng Hu, Sizhe Liu, Changjiang Gao, Xin Huang, Xue Han, Junlan Feng, Chao Deng, and Shujian Huang. 2025. Large language models are cross-lingual knowledge-free reasoners. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1525–1542, Albuquerque, New Mexico. Association for Computational Linguistics.

Zheng Wei Lim, Alham Fikri Aji, and Trevor Cohn. 2025. Language-specific latent process hinders cross-lingual performance.

Xin Liu, Qiyang Song, Qihang Zhou, Haichao Du, Shaowen Xu, Wenbo Jiang, Weijuan Zhang, and Xiaoqi Jia. 2026. Focusing on language: Revealing and exploiting language attention heads in multilingual large language models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(38):32195–32203.

Yanan Long. 2025. Towards multilingual mechanistic interpretability. In *NeurIPS 2025 Workshop on Evaluating the Evolving LLM Lifecycle: Benchmarks, Emergent Abilities, and Scaling*.

Weicheng Ma, Kai Zhang, Renze Lou, Lili Wang, and Soroush Vosoughi. 2021. Contributions of transformer attention heads in multi- and cross-lingual tasks. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1956–1966, Online. Association for Computational Linguistics.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*, volume 35, pages 17359–17372.

Paul Michel, Omer Levy, and Graham Neubig. 2019. Are sixteen heads really better than one? In *Advances in Neural Information Processing Systems*, volume 32.

Aaquib Syed, Can Rager, and Arthur Conmy. 2024. Attribution patching outperforms automated circuit discovery. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 407–416, Miami, Florida, US. Association for Computational Linguistics.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. 2019. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5797–5808, Florence, Italy. Association for Computational Linguistics.

Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. 2023. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. In *The Eleventh International Conference on Learning Representations*.

Mingyang Wang, Heike Adel, Lukas Lange, Yihong Liu, Ercong Nie, Jannik Strötgen, and Hinrich Schuetze. 2025. Lost in multilinguality: Dissecting cross-lingual factual inconsistency in transformer language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5075–5094, Vienna, Austria. Association for Computational Linguistics.

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 (Volume 1: Long Papers)*, pages 15366–15394, Bangkok, Thailand. Association for Computational Linguistics.

Weihao Xuan, Rui Yang, Heli Qi, Qingcheng Zeng, Yunze Xiao, Aosong Feng, Dairui Liu, Yun Xing, Junjue Wang, Fan Gao, Jinghui Lu, Yuang Jiang, Huitao Li, Xin Li, Kunyu Yu, Ruihai Dong, Shangding Gu, Yuekang Li, Xiaofei Xie, Felix Juefei-Xu, Foutse Khomh, Osamu Yoshie, Qingyu Chen, Douglas Teodoro, Nan Liu, Randy Goebel, Lei Ma, Edison Marrese-Taylor, Shijian Lu, Yusuke Iwasawa, Yutaka Matsuo, and Irene Li. 2025. MMLU-ProX: A multilingual benchmark for advanced large language model evaluation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 1513–1532, Suzhou, China. Association for Computational Linguistics.

An Yang et al. 2025. Qwen3 technical report.

Chengzhi Zhong, Qianying Liu, Fei Cheng, Junfeng Jiang, Zhen Wan, Chenhui Chu, Yugo Murawaki, and Sadao Kurohashi. 2025. What language do non-English-centric large language models think in? In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 26333–26346, Vienna, Austria. Association for Computational Linguistics.

A Prompting, answer scoring, and answer-order perturbations

Benchmark prompts. For Belebele, we use:

```
Passage: {passage}
Question: {question}
Options:
{options}
Answer:
```

For MMLU-ProX, we use:

```
Question: {question}
Options:
{options}
Answer:
```

The `{options}` field prints one labeled option per line, e.g., `A. <option text>`, `B. <option text>`, and so on.

Answer extraction. We evaluate behavioral correctness by greedy generation from the answer prefix. We generate up to two tokens with temperature 0, decode the output, and extract the first standalone option label in the valid label set after stripping leading whitespace. A prediction is correct if this label matches the gold label; outputs with no valid label are marked incorrect.

Option-label scores. For the intervention analyses, the option score $z_{i,c}$ is the model's log-probability for the option-label string following the answer prefix and separator. Concretely, for label c we score the string consisting of `answer_sep` followed by c after the prompt ending in `Answer:`. These label scores are used to compute the correct-vs-average-incorrect margin in Section 3.1.

Answer-order perturbations. For each retained question, we generate four answer-order permutations. The permutations preserve answer contents and the identity of the correct answer, but vary the label assigned to each option. Each permutation is patched independently, and the resulting per-group importance tensors are flattened and averaged into the usage signature v_i^ℓ. The retained question set is fixed across permutations.

B Attention-group patching details

The models use grouped-query attention, where multiple query heads share a smaller set of key-value heads. We define an attention group as the slice of the pre-output-projection attention result associated with all query heads that share one key-value head. For a layer with hidden size d and G key-value groups, we view the pre-output-projection attention result as a concatenation of G group slices. Patching a group replaces only one such slice at the final token position; the modified concatenation is then passed through the usual attention output projection.

Table 6 reports the intervention dimensionality for each model. The signature dimension is $H = L \times G$, where L is the number of layers and G is the number of attention groups per layer.

Model	L	n_q	G	H
Qwen3-8B	36	32	8	288
Llama-3.1-8B	32	32	8	256

Table 6: Attention-group intervention dimensions. L is the number of layers, n_q is the number of query heads, G is the number of key-value groups per layer, and $H = L \times G$ is the signature dimension.

Donor means are computed separately for each model, dataset, and language using the retained multilingual-correct subset for that model and dataset. Donor means are not shared across models or datasets. Answer-order perturbations are used when computing intervention scores, but the retained set itself is defined using the original answer ordering.

For each retained question-language pair, we:

1. run the unpatched model and cache final-token attention-group outputs at every layer;
2. compute the language-matched donor mean $d_{\ell,l,g}$ for each layer and group;
3. rerun the model once for each candidate group (l, g);
4. replace only that group's final-token output with $d_{\ell,l,g}$;
5. compute the change in the correct-vs-average-incorrect margin.

C Artifact use and licensing

We use Qwen3-8B, Llama-3.1-8B, Belebele, and MMLU-ProX for research evaluation and mechanistic analysis, and cite their creators in the main text. We do not redistribute model weights or benchmark data. We follow the publicly released licenses and terms of use for each artifact.

We created the latent-country dataset for this work as a controlled research artifact for studying intermediate computation and cross-lingual patching. The dataset contains country-related factual entries, including capitals, calling codes, cities, landmarks, and universities. It does not contain personal or sensitive user data. The English knowledge base was manually fact-checked, and translations into Spanish, German, Russian, and Arabic were reviewed by native speakers. If released, the dataset will be distributed with a license and usage statement.

D Latent-country dataset construction and scoring

The latent-country dataset covers 25 countries. For each country, we record one capital, one international calling code, five cities, five landmarks, and five universities. The English knowledge base was constructed semi-automatically with assistance from GPT-5 and was then manually fact-checked. The English data was subsequently translated into Spanish, German, Russian, and Arabic, with each translation verified by a native speaker. The resulting files are aligned by country while using language-appropriate surface forms.

Prompts are generated from these aligned files. The normal prompt always uses

city→capital. The alternative prompt is drawn from either landmark→country code or university→country code. The hidden country never appears explicitly in either prompt. Each normal and alternative prompt contains 20 in-context examples, and each experimental condition is evaluated on 100 sampled normal-alternative prompt pairs. For cross-lingual conditions, the normal and alternative prompts are drawn from different language files; $\ell' \rightarrow \ell$ denotes an alternative prompt in language ℓ' and a normal prompt in language ℓ.

Latent-country scoring. For each normal-alternative prompt pair, the target is the type-corrected alternative answer: the answer associated with the alternative prompt's hidden country, expressed in the target type and language of the normal prompt. We evaluate promotion of this target using vocabulary ranks at the first answer position. Let t_i be the first tokenizer token of the type-corrected alternative answer when preceded by the answer separator. We rank t_i among all tokens in the model vocabulary using its logit at the first answer position. T1 is the percentage of examples in which t_i is ranked first, T10 is the percentage in which t_i is ranked among the top 10, and rank is the vocabulary rank of t_i. For multi-token surface forms, these metrics measure promotion of the first answer token rather than the full answer string.

E Per-language-pair signature similarity

Table 7 reports aligned-minus-null gaps for each language pair. All gaps are positive; for all rows and both metrics, the one-sided permutation test attains the minimum attainable p-value with $B = 500$ shuffled null draws, $p = 1/(B+1) = 1/501 \approx 0.002$.

F Layerwise variance decomposition

Figures 4 and 5 show the layerwise SS-weighted variance decomposition of the rank-normalized importance tensor for each model, averaged over Belebele and MMLU-ProX. Because Qwen3-8B and Llama-3.1-8B have different numbers of layers, we aggregate layerwise decompositions within each model rather than averaging raw layer indices across models.

Dataset	Model	Pair	Δ_ρ	$\Delta_{\mathrm{J@10}}$
Belebele	Qwen3-8B	en–es	0.221	0.111
Belebele	Qwen3-8B	en–de	0.227	0.112
Belebele	Qwen3-8B	en–ru	0.207	0.109
Belebele	Qwen3-8B	en–ar	0.143	0.073
Belebele	Qwen3-8B	es–de	0.319	0.145
Belebele	Qwen3-8B	es–ru	0.284	0.135
Belebele	Qwen3-8B	es–ar	0.210	0.104
Belebele	Qwen3-8B	de–ru	0.296	0.142
Belebele	Qwen3-8B	de–ar	0.203	0.096
Belebele	Qwen3-8B	ru–ar	0.213	0.101
Belebele	Llama-3.1-8B	en–es	0.302	0.183
Belebele	Llama-3.1-8B	en–de	0.262	0.158
Belebele	Llama-3.1-8B	en–ru	0.262	0.169
Belebele	Llama-3.1-8B	en–ar	0.174	0.112
Belebele	Llama-3.1-8B	es–de	0.305	0.179
Belebele	Llama-3.1-8B	es–ru	0.309	0.201
Belebele	Llama-3.1-8B	es–ar	0.215	0.133
Belebele	Llama-3.1-8B	de–ru	0.294	0.179
Belebele	Llama-3.1-8B	de–ar	0.207	0.136
Belebele	Llama-3.1-8B	ru–ar	0.234	0.139
MMLU-ProX	Qwen3-8B	en–es	0.263	0.126
MMLU-ProX	Qwen3-8B	en–de	0.240	0.119
MMLU-ProX	Qwen3-8B	en–ru	0.236	0.116
MMLU-ProX	Qwen3-8B	en–ar	0.179	0.100
MMLU-ProX	Qwen3-8B	es–de	0.398	0.174
MMLU-ProX	Qwen3-8B	es–ru	0.347	0.156
MMLU-ProX	Qwen3-8B	es–ar	0.297	0.134
MMLU-ProX	Qwen3-8B	de–ru	0.350	0.167
MMLU-ProX	Qwen3-8B	de–ar	0.285	0.135
MMLU-ProX	Qwen3-8B	ru–ar	0.317	0.151
MMLU-ProX	Llama-3.1-8B	en–es	0.452	0.269
MMLU-ProX	Llama-3.1-8B	en–de	0.411	0.232
MMLU-ProX	Llama-3.1-8B	en–ru	0.399	0.232
MMLU-ProX	Llama-3.1-8B	en–ar	0.326	0.190
MMLU-ProX	Llama-3.1-8B	es–de	0.489	0.284
MMLU-ProX	Llama-3.1-8B	es–ru	0.479	0.281
MMLU-ProX	Llama-3.1-8B	es–ar	0.414	0.228
MMLU-ProX	Llama-3.1-8B	de–ru	0.466	0.267
MMLU-ProX	Llama-3.1-8B	de–ar	0.400	0.223
MMLU-ProX	Llama-3.1-8B	ru–ar	0.430	0.241

Table 7: Per-language-pair aligned-minus-null gaps for signature similarity. Δ_ρ is the gap for Spearman similarity, and $\Delta_{\mathrm{J@10}}$ is the gap for Jaccard@10.

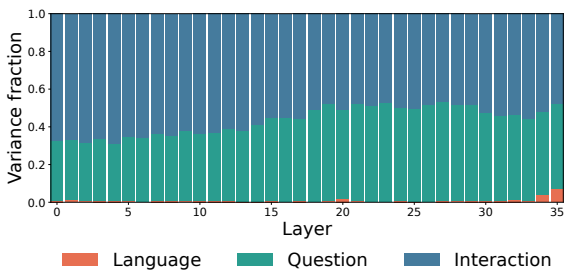

Figure 4: Layerwise SS-weighted variance decomposition for Qwen3-8B, averaged over Belebele and MMLU-ProX. Bars show the fraction of within-layer variation attributed to language identity, question identity, and language-question interaction; fractions sum to one within each layer.

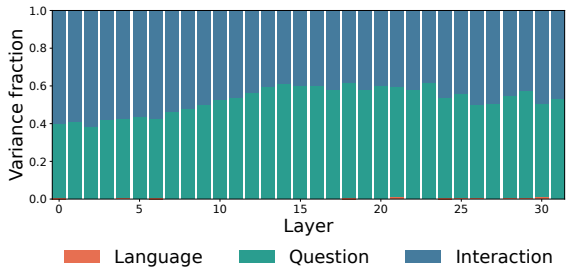

Figure 5: Layerwise SS-weighted variance decomposition for Llama-3.1-8B, averaged over Belebele and MMLU-ProX. Bars show the fraction of within-layer variation attributed to language identity, question identity, and language-question interaction; fractions sum to one within each layer.

G Per-condition latent-country results

This appendix reports the per-condition vocabulary-rank improvements aggregated in Table 5. Each row of the heatmaps corresponds to one experimental condition, specified by the language pair $\ell' \rightarrow \ell$ and the alternative-prompt family: `landmark→country code` or `university→country code`. The normal-prompt task is always `city→capital`. The columns *Base T1* and *Patch T1* are the baseline and patched percentages of examples where the first token of the type-corrected alternative answer is ranked first in the full model vocabulary. *Delta T1* is their difference, and *Patch T10* is the percentage of examples where that token is ranked in the top 10 after patching.

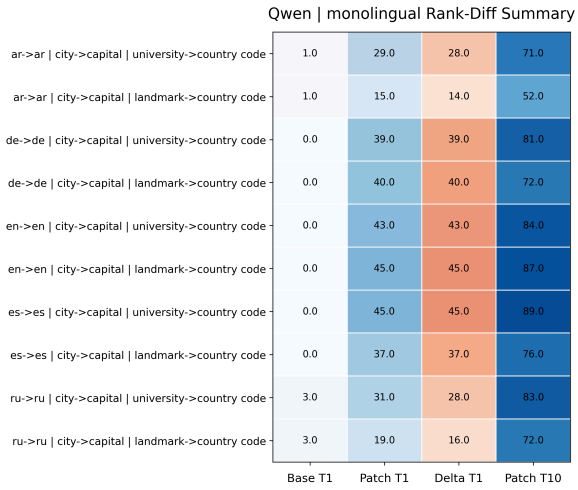

Figure 6: Qwen3-8B monolingual per-condition vocabulary-rank summary.

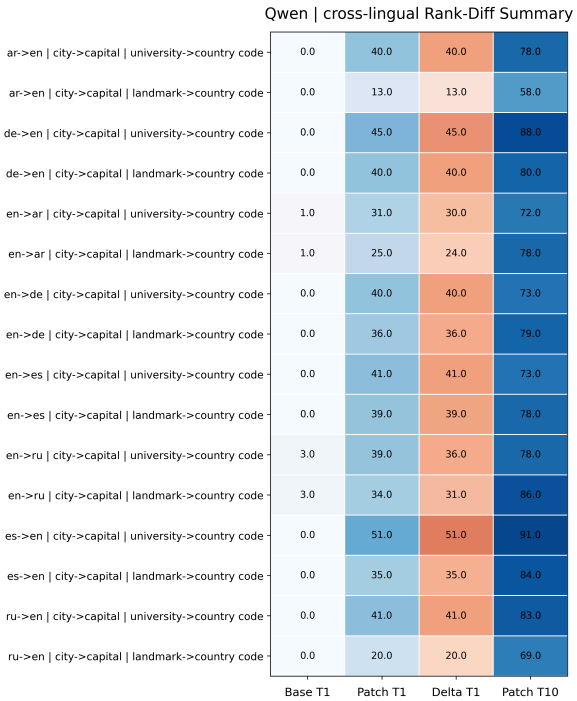

Figure 7: Qwen3-8B cross-lingual per-condition vocabulary-rank summary.

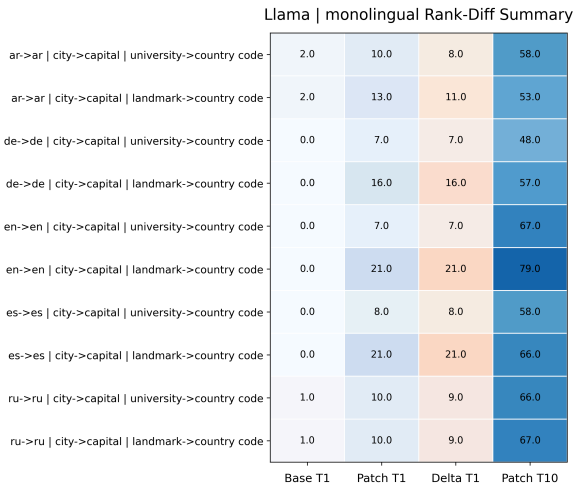

Figure 8: Llama-3.1-8B monolingual per-condition vocabulary-rank summary.

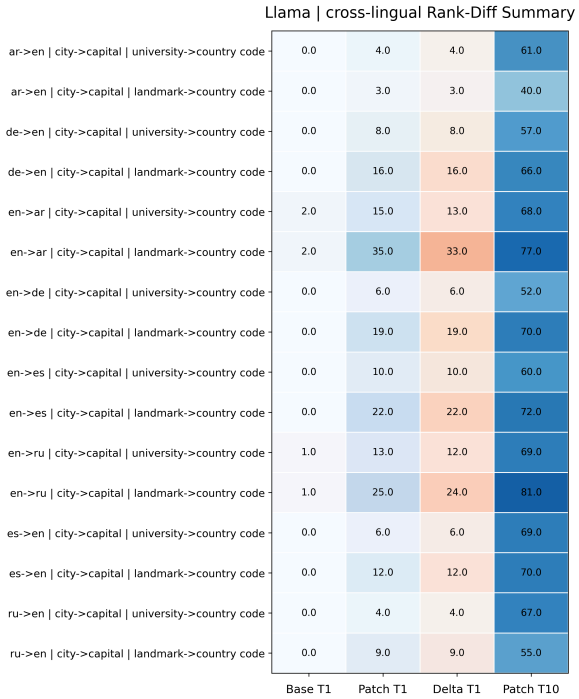

Figure 9: Llama-3.1-8B cross-lingual per-condition vocabulary-rank summary.