Shared Circuits for Shared Grammar: Tracing Subject-Verb Agreement Across Languages

Isabella Gidi
Harvard University
isabellagidi@college.harvard.edu

Antonio Almudévar
University of Zaragoza

Core Francisco Park
Harvard University

Naomi Saphra
Boston University

Ricard Marxer
Univ Toulon, Aix Marseille Univ, CNRS, LIS & ILLS

Abstract

Multilingual large language models often generalize across languages, and prior work suggests that their internal mechanisms can overlap cross-lingually. It remains unclear, however, when such sharing emerges and whether it varies with the overt realization of the same grammatical operation. We investigate this question for present-tense subject-verb agreement, a morphosyntactic process that varies substantially across languages and is only weakly expressed in English. Using activation patching and attention analysis across 29 languages and five open-source model families, we identify the attention heads causally implicated in agreement and compare these head-level signatures across languages. We find that languages with overt person/number inflection exhibit more similar agreement circuitry than non-conjugating languages, with the strongest sharing appearing when the analysis isolates recovery of the inflectional contrast itself. English provides an informative bridge case, becoming more similar to conjugating languages precisely in contexts where overt agreement is required. Finally, many implicated heads display similar attention patterns across languages, suggesting that cross-lingual overlap reflects shared functional roles as well as shared localization. Together, these results indicate that multilingual LLMs reuse partially shared computational structure for morphosyntactic agreement rather than relying on fully separate language-specific solutions.

1 Introduction

Multilingual large language models (LLMs) generalize across a wide range of languages, yet we still do not understand the extent to which their internal computations are shared. In particular, it remains unclear when multilingual models rely on language-specific mechanisms and when they reuse common internal circuitry. This distinction matters because it shapes what kinds of cross-lingual generalization we should expect inside the model: if the same computation is implemented by shared circuitry, then interpretability findings, causal interventions, and alignment methods may transfer across languages; if it is implemented differently, then analyses and interventions developed in one language may fail to generalize to others.

This question is especially important for morphology and syntax. Languages differ in these domains in ways that may not map straightforwardly onto a single language-agnostic internal representation, especially when a high-resource language lacks the relevant feature altogether. This raises a central mechanistic question: when multilingual models process the same morphosyntactic phenomenon across typologically different languages, do they recruit *shared* internal circuitry or *independent* language-specific solutions?

Prior work has established that cross-lingual circuit overlap can occur in focused case studies; the unresolved question is what determines when multilingual models reuse the same circuitry and when they recruit language-specific mechanisms.

We investigate this question through the lens of subject-verb agreement (Appendix A.1). Present-tense verb conjugation provides a particularly informative testbed because the same grammatical dimensions are realized differently across languages: person and number are overtly marked in many languages, absent in languages such as Chinese, and only partially marked in English. Here, we use *conjugation* to mean changes in a verb's surface form as a function of the subject's grammatical person and number. Using activation patching and attention analysis across 29 languages and six checkpoints from five model families, we test whether agreement-circuit similarity varies systematically with the requirement to overtly realize this grammatical operation, rather than asking only whether cross-lingual overlap can be detected. We further distinguish circuitry that broadly supports recovery of the context-appropriate target form from circuitry that specifically restores the inflectional contrast in controlled minimal pairs.

Our main findings are as follows:

1. **Conjugation circuitry is partially shared across languages:** Among languages with overt subject-verb agreement, we find substantial overlap in the attention heads implicated in conjugation.
2. **Shared circuitry is strongest for inflection-specific recovery:** We distinguish between broader recovery of the context-appropriate target form and stricter recovery of the inflectional contrast itself. Similarity is highest under the minimal-pair metric, suggesting that multilingual models share circuitry most strongly for overt inflectional agreement, while broader target-form recovery is only partially shared and more variable across languages.
3. **Shared circuitry tracks shared grammatical structure:** Languages with little or no person-based verb inflection show substantially lower similarity to these head-level signatures. English provides an informative bridge case: it differs from strongly conjugating languages overall, but its third-person singular cases align more closely with the shared conjugation circuitry than its non-inflecting cases do.
4. **Overlapping heads play similar functional roles:** Across conjugating languages, shared heads show similar attention patterns, suggesting that cross-lingual overlap reflects shared information routing rather than only shared localization.

2 Related Work

Shared and language-specific multilingual mechanisms. A substantial literature suggests that multilingual models often rely on partially shared cross-lingual structure, especially for semantics, factual recall, and safety-related behavior (Wendler et al., 2024; Dumas et al., 2025; Wang et al., 2025). Work on grammar and morphosyntax likewise finds evidence of shared organization, including overlapping grammatical subspaces, shared neuron sets, and agreement-related components (Stanczak et al., 2022; Mueller et al., 2022; Ferrando & Costa-jussà, 2024). At the same time, other studies identify language-specific neurons or features and show that cross-lingual transfer and interventions do not generalize uniformly across languages (Tang et al., 2024; Deng et al., 2025; Zhang et al., 2025). Taken together, this literature suggests that the key question is not whether multilingual computation is globally shared or language-specific, but *when* each pattern arises. This question is especially important for morphosyntax, where languages may share broad meaning while differing in whether they overtly realize a grammatical operation at all. We therefore study present-tense subject-verb agreement as a test case for when multilingual models reuse shared internal mechanisms and when they diverge.

From shared representations to shared circuitry. Cross-lingual structure has been studied with probing, representational analyses, neuron and feature methods, and causal interventions (Mikhailov et al., 2021; Brinkmann et al., 2025; Jing et al., 2025; Mueller et al., 2022). Our work is closest to causal studies of agreement and cross-lingual circuit similarity (Mueller et al., 2022; Ferrando & Costa-jussà, 2024; Zhang et al., 2025), but differs in three ways most relevant to our question. First, we use tightly controlled multilingual minimal pairs that isolate the inflectional contrast. Second, we compare a broader and more typologically

varied set of languages, including languages with rich agreement, partial agreement, and no overt agreement in the tested contrasts. Third, we use two complementary recovery metrics: target-logit recovery, which measures broader recovery of the context-appropriate form under a subject change, and minimal-pair recovery, which measures recovery of the inflectional contrast itself. This allows us to ask separately whether multilingual models share mechanisms for broader subject-conditioned prediction and whether they share mechanisms for implementing overt agreement morphology.

3 Experimental Setup

Models evaluated. We investigate conjugation circuitry across five open-source model families: BLOOM (BigScience Workshop et al., 2022), Gemma (Gemma Team, 2024), Llama (Touvron et al., 2023; Grattafiori et al., 2024), Mistral (Jiang et al., 2023), and Qwen (Yang et al., 2024). The full model list is provided in Appendix A.2, with additional analyses of model size and instruction tuning in Appendix D.

Verb dataset. We construct our candidate verb inventory from structured lexical and morphological resources. For 33 languages, we extract verb infinitives and their corresponding present-tense inflected forms from UniMorph (Batsuren et al., 2022), which provides isolated morphological paradigms. We additionally use the Chinese Verb Semantic Feature Dataset (Deng et al., 2024) to obtain verb lemmas for Chinese, providing a high-resource non-conjugating comparison language. At this stage, these resources provide only the lexical items and their relevant forms; the experimental prompts are subsequently constructed from these entries using the controlled template described below. Across languages, the number of distinct verb infinitives with corresponding conjugated forms ranges from 56 to 23,782, with an average of 1,750.74. In total, our initial inventory spans 34 languages.

Inflection contrasts. To keep the analysis tractable while covering the main morphosyntactic dimensions of interest, we evaluate a representative subset of bidirectional person and number contrasts. Specifically, we consider shifts in person while holding number constant (1sg $\leftrightarrow$ 2sg, 1sg $\leftrightarrow$ 3sg, and 1pl $\leftrightarrow$ 3pl), as well as a shift in number while holding person constant (3sg $\leftrightarrow$ 3pl). Of these four pairs, English has an inflectional change on two (1sg $\leftrightarrow$ 3sg, 3sg $\leftrightarrow$ 3pl).

Prompt template. For each language, we generate zero-shot prompts designed to elicit the target inflection. By using a common prompt structure, we largely isolate differences in agreement-related circuitry. The prompts follow a common template of the form: *"Conjugation of the verb [infinitive] in present tense: [pronoun] [conjugation]"*, adapted into each target language (Adaptations in Appendix A.3). This deliberately structured elicitation format preserves positional alignment between clean and corrupted prompts, which is important for controlled activation patching, but trades naturalism for experimental control. In preliminary experiments across models, this phrasing maximized the probability assigned to the correct conjugation.

Tokenization filtering. To construct tightly controlled minimal pairs in conjugating languages, we apply a model-specific tokenization filter and retain only verbs whose clean and corrupted conjugated forms differ only in the final token. This preserves alignment in the prefix sequence and isolates the person or number contrast to the token the model must predict. Examples of retained and excluded cases are provided in Appendix A.4.

Behavioral filtering. Because our goal is to analyze the mechanisms involved in successful agreement computation, we also remove examples where the model does not produce the desired verb inflection. Following prior work on isolating task-relevant circuitry (e.g., Todd et al., 2024), this restriction focuses the analysis on successful realizations of grammatical agreement rather than heterogeneous error behavior. We thus interpret our analyses as characterizing the circuitry used in successful conjugation behavior rather than overall language competence.

Language-model coverage. Official language support varies across model families, but many models achieve sufficient accuracy on languages beyond those explicitly documented in their training descriptions. We include a language-model combination in our analysis only if at least 50 prompts remain after tokenization and behavioral filtering. Starting from 34 languages, this yields coverage for 29 languages in the main analyses. These retained languages include 24 languages with overt agreement patterns, 4 languages with no overt agreement in the tested contrasts, and English as an intermediate case, since it marks agreement only in the third-person singular. The resulting language-model coverage and full language inventory are provided in Appendix A.5.

Final dataset. After tokenization and behavioral filtering, the number of surviving prompts ranges from 96 to 2,400 across the retained language-model combinations, with an average of 1,509.59. From each retained set, we evaluate 50 prompts for activation patching.

4 Analytic Methods

To identify and characterize the internal circuitry involved in subject-verb agreement, we use attention-output patching to localize causally important attention heads and attention analysis to characterize their behavior.

4.1 Attention-output Patching

We use attention-output patching (Vig et al., 2020; Zhang & Nanda, 2024) to identify the attention heads causally important for predicting the target verb form under subject manipulations (Figure 1).

Minimal pairs. Minimal pairs are constructed as described in Section 3. For each pair, we define a *clean* prompt corresponding to the target inflection and a *corrupted* prompt corresponding to the contrastive inflection. For example, when evaluating a 1sg $\rightarrow$ 3sg shift, the third-person form is the clean prompt and the first-person form is the corrupted prompt.

Head-level patching. During a forward pass on the corrupted prompts, we replace the output of attention head h in layer ℓ with its cached activation from the clean run at the corresponding sequence position. Let $\mathbf{L}_{\text{patch}}^{(\ell,h,i)}$ denote the resulting final-layer logit vector for prompt i, let $\mathbf{L}_{\text{corr}}^{(i)}$ denote the unpatched corrupted logits, and let $y_{\text{clean}}^{(i)}$ denote the clean target token.

Recovery metrics. We use two complementary recovery metrics: target-logit recovery and minimal-pair recovery. We use target-logit recovery for analyses that include non-conjugating languages, and minimal-pair recovery for analyses restricted to conjugating languages. For both metrics, larger absolute values indicate stronger effects overall, with positive values reflecting promotion of the clean target token and negative values reflecting inhibition. Unless otherwise noted, each metric is computed over 50 prompts per conjugation pair, evaluated in both patching directions (e.g., 1sg $\rightarrow$ 3sg and 3sg $\rightarrow$ 1sg), and aggregated into layer-head heatmaps.

Target-logit recovery. Target-logit recovery measures the increase in the clean target token's logit relative to the corrupted baseline:

$$M_{\text{target}}(\ell, h) = \frac{1}{n} \sum_{i=1}^{n} \left(\left[\mathbf{L}_{\text{patch}}^{(\ell,h,i)} \right]_{y_{\text{clean}}^{(i)}} - \left[\mathbf{L}_{\text{corr}}^{(i)} \right]_{y_{\text{clean}}^{(i)}} \right).$$

Because this metric does not require a contrast between two different inflected forms, it can be applied both to conjugating languages and to languages or conjugation pairs without overt present-tense agreement inflection. In conjugating languages, it captures how much patching restores the clean inflected form relative to the corrupted prompt. In languages

or conjugation pairs without overt agreement inflection, the clean and corrupted prompts may share the same final target token, but the subject manipulation can still change the logit assigned to that token. In those cases, the metric asks whether patching restores the clean-context preference for the shared target form. We therefore interpret the identified heads as supporting successful prediction of the target form in context, not always as encoding overt inflectional contrast.

Minimal-pair recovery. Minimal-pair recovery measures how much patching restores the clean-versus-corrupted logit contrast. For each example i, let $y^{(i)}_{\text{clean}}$ and $y^{(i)}_{\text{corr}}$ denote the competing final inflection tokens, and define the contrast for a final-layer logit vector $\mathbf{L}^{(i)}$ as

$$\Delta^{(i)}(\mathbf{L}^{(i)}) = L^{(i)}_{y^{(i)}_{\text{clean}}} - L^{(i)}_{y^{(i)}_{\text{corr}}}.$$

We write Δ_{clean}, Δ_{corr}, and $\Delta(\mathbf{L}^{(\ell,h)}_{\text{patch}})$ for this contrast averaged across examples under the unpatched clean run, the unpatched corrupted run, and the run in which attention head h in layer ℓ is patched, respectively. Minimal-pair recovery is then defined as

$$M_{\text{pair}}(\ell, h) = \frac{\Delta(\mathbf{L}^{(\ell,h)}_{\text{patch}}) - \Delta_{\text{corr}}}{\Delta_{\text{clean}} - \Delta_{\text{corr}} + \epsilon},$$

where ϵ is a small constant added for numerical stability. This metric is applicable only when the clean and corrupted prompts have different final inflection tokens and therefore provides a targeted measure of how strongly patching restores the preference for the context-appropriate inflected form over its contrastive alternative.

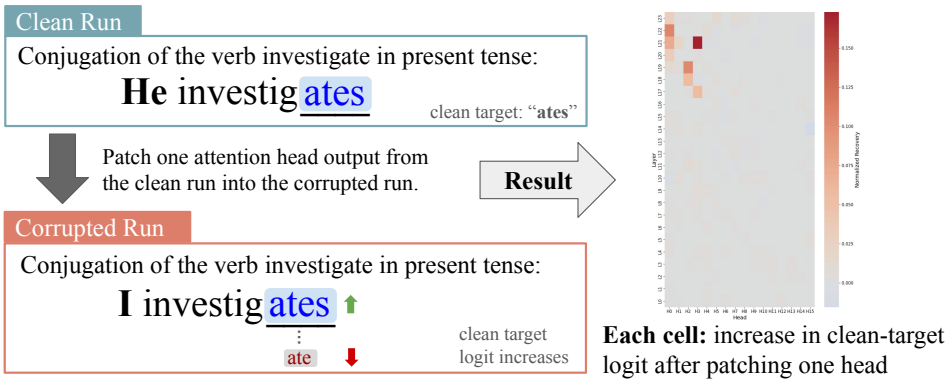


Figure 1: **Example of activation patching.** The clean and corrupted prompts differ only in the subject pronoun and target inflection, with the contrast isolated to the final token. For each layer-head pair, we patch that head's output from the clean run into the corrupted run and measure the resulting increase in the clean-target logit. The resulting heatmap summarizes the causal contribution of each head.

Cross-lingual comparison. To test whether languages rely on similarly localized heads for the conjugation task, we compare activation-patching heatmaps across languages, treating these heatmaps as signatures of the circuitry implicated in agreement. For each language, we flatten the layer–head heatmap into a vector and compute Pearson correlation between language pairs within the same model, conjugation pair, and patching direction. Because our goal is to compare the locations of task-relevant heads rather than the direction of their effects, our primary analyses use the absolute values of the heatmaps before flattening. Under this measure, high similarity indicates that two languages assign relatively high importance to the same layer-head locations. Across all boxplots, n denotes the number of observations contributing to the plotted distribution. We report alternative similarity measures, including Spearman correlation and top-k head overlap, in Appendix C.

Cross-model comparison. We never compare attention-head coordinates or patch activations directly across different architectures or checkpoints. For each checkpoint, we compute cross-lingual similarity between language heatmaps matched by conjugation contrast and patching direction. We then report these within-model similarities separately by checkpoint or aggregate their scalar similarity values across checkpoints for summary analyses.

4.2 Attention Analysis

Activation patching identifies which heads are causally important for agreement, but not how those heads behave once implicated. To characterize their functional behavior, we analyze attention patterns (Li et al., 2016), focusing on attention emitted from the final prefix position, i.e., when the model is predicting the final inflection token.

Top-head selection. For each model, language, conjugation pair, patching direction, and recovery metric, we rank attention heads by the absolute magnitude of their patching score and retain the top $k = 20$ heads. We compute the attention-role statistics described below separately for each selected head. For cross-lingual comparisons, we restrict the analysis to overlapping heads: heads that appear among the top 20 for both languages within the same model, conjugation pair, patching direction, and metric. We then compare the corresponding attention-role vectors for these shared heads across the two languages.

Attention-role summaries. To interpret these patterns, we partition each prompt prefix into linguistically relevant categories using tokenizer-specific alignments derived from the prompt annotations. For each example, we measure the fraction of attention mass from the final prefix position directed to the subject pronoun, the explicit infinitive span, and all remaining tokens. These role-mass values are averaged first across examples and then across patching directions within each language-pair condition. We use these summaries to compare conjugating and non-conjugating languages, as well as English non-inflecting and third-person singular contexts.

Cross-lingual comparison. To assess cross-lingual reuse beyond role-mass summaries, we compare the attention-role vectors associated with heads identified across languages. In the main analysis, we evaluate whether overlapping top-ranked heads exhibit similar role profiles across languages. Together, these analyses test whether heads that are similarly localized by activation patching also exhibit similar functional behavior across languages.

5 Results

Results roadmap. We organize our results by moving from structural localization to mechanistic behavior. To localize the relevant circuitry, we rely on the distinction between our two recovery metrics. The *minimal-pair* metric provides a sharp test of shared *inflection-specific* circuitry among conjugating languages, whereas the broader *target-logit* metric allows us to test whether subject-conditioned prediction generalizes to non-conjugating languages. Finally, after identifying these cross-lingual networks, we analyze their fine-grained attention patterns to determine whether these structurally shared heads execute the same information-routing mechanisms across languages.

5.1 Inflection-specific sharing is strongest under the minimal-pair metric

We first investigate whether languages with overt present-tense agreement recruit similar head-level circuitry, and how the choice of recovery metric influences this shared structure.

Conjugating languages share robust inflectional circuitry. Under the minimal-pair metric, pairwise similarities among conjugating languages are strongly positive across model families (Figure 2, left). This indicates that the head-level structures implicated by activation patching are broadly shared across inflecting languages. While the exact strength and dispersion of this pattern vary by model architecture, language-level similarity matrices

show that the effect is broadly distributed across languages rather than being restricted to a small set of closely related language pairs (Appendix B.1). This positive cross-lingual similarity also persists across all individual conjugation contrasts and patching directions (Appendix B.2).

Sharing peaks during strict inflection-specific recovery. Comparing the two metrics reveals a critical functional distinction (Figure 2, right). Similarity among conjugating languages is higher and less variable under the minimal-pair metric than under the target-logit metric. This demonstrates that multilingual models share circuitry most intensely when restoring the strict inflectional contrast itself. Conversely, the broader prediction of the context-appropriate target form (measured by target-logit) exhibits only partial cross-lingual sharing, indicating it is more sensitive to language-specific computation.

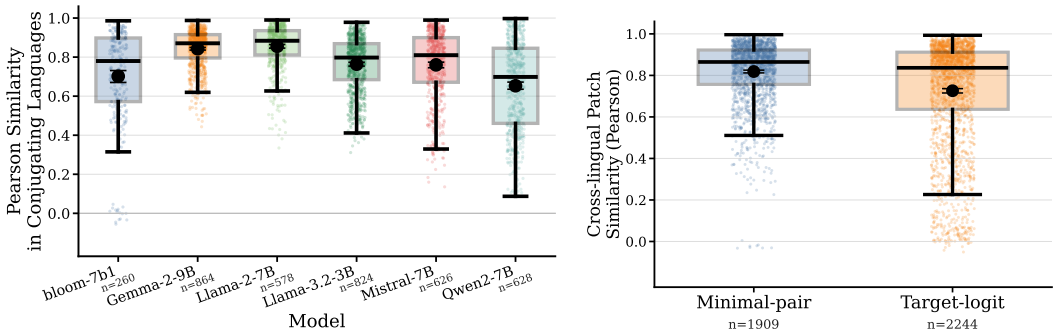


Figure 2: **Inflection-specific sharing is stronger under the minimal-pair metric. Left:** Pairwise similarity among conjugating languages under the minimal-pair metric, shown separately by model checkpoint. Similarity is broadly positive across families, indicating shared inflection-related circuitry across languages with overt agreement. **Right:** Pairwise similarity among conjugating languages under the minimal-pair and target-logit metrics. Similarity is higher and less variable under the minimal-pair metric, indicating that cross-lingual sharing is strongest when the analysis isolates recovery of the inflectional contrast itself.

Summary. Collectively, these results support a graded view of cross-lingual structural alignment. Rather than relying on entirely disjoint, language-specific mechanisms for present-tense agreement, models reuse a shared foundational circuitry. Crucially, this sharing is most pronounced in the specific circuits dedicated to resolving the inflectional contrast, whereas broader target-form prediction introduces more language-specific variance. Auxiliary analyses regarding model size, language-specific task capability, instruction tuning, and BLOOM training checkpoints are detailed in Appendix D.

5.2 Broader target-form recovery is less shared, but still separates conjugating from non-conjugating languages

We next ask whether the broader *target-logit* metric still reveals linguistically structured sharing. Because this metric can be applied even when a language lacks overt present-tense agreement, it allows us to test whether subject-conditioned target recovery generalizes to non-conjugating languages. English is treated separately, as it occupies an intermediate position by marking agreement only in the third-person singular.

Language type dictates shared circuitry. Figure 3 (left) demonstrates a clear separation by language type. Comparisons between two conjugating languages yield much higher similarity than comparisons across the conjugating/non-conjugating boundary. This indicates that the recovered head-level signatures are not generic; they are strongly shaped by whether the language overtly realizes the morphosyntactic feature. Furthermore, similarity among non-conjugating languages remains relatively low, proving that the metric is not simply recovering a universal, baseline prediction pattern for all non-conjugating languages.

This result complements Section 5.1: while the minimal-pair metric isolates strictly *inflection-specific* shared circuitry, the broader target-logit metric still sharply delineates languages with overt agreement from those without. Despite its higher variance, it reveals a clear, cross-lingual grammatical organization.

English dynamically shifts based on grammatical context. English provides an exceptionally informative test case. Because it marks present-tense agreement only in third-person singular (3sg) contexts, it allows us to isolate whether cross-lingual similarity is driven by static language identity or by the contextual demand for an agreement computation. As shown in Figure 3 (right), English exhibits a consistent upward shift from non-inflecting to 3sg-inflecting contexts. It becomes markedly more similar to conjugating languages exactly when overt agreement is required.

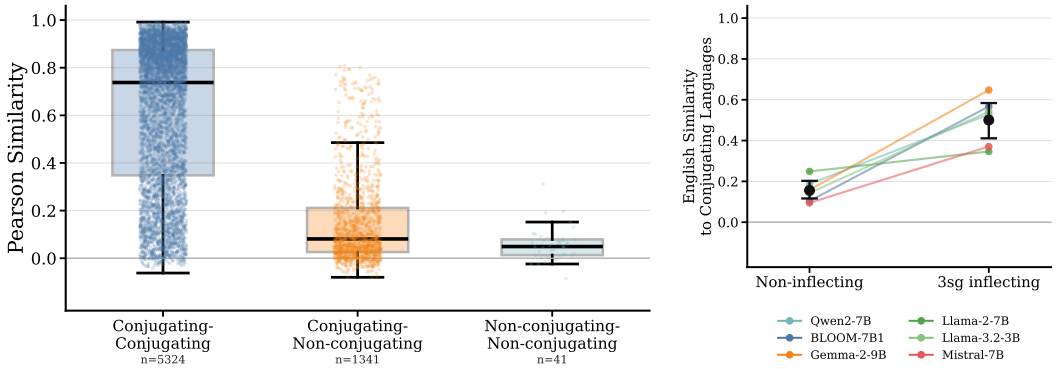


Figure 3: **Broader target-form recovery still tracks overt agreement structure. Left:** Pairwise similarity between activation-patching heatmaps grouped by language comparison type under the target-logit metric. Each point is a comparison between two heatmaps matched for model, conjugation pair, and patching direction. Conjugating–conjugating comparisons are substantially more similar than conjugating–non-conjugating comparisons, showing that broader target-form recovery remains structured by whether the language overtly marks agreement. **Right:** English similarity to conjugating languages in non-inflecting versus 3sg-inflecting contexts. Each colored line corresponds to one model and connects its mean English–to–conjugating similarity across the two condition types. Black points and error bars show the across-model mean and bootstrap confidence interval for each condition. English becomes more similar to conjugating languages precisely in the contexts where overt agreement is required.

Summary. Taken together, these results demonstrate that while broader target-form recovery is less uniformly shared than inflection-specific recovery, it remains strictly organized by grammatical structure. Languages with overt agreement form a distinct similarity cluster, and English shifts toward this cluster precisely when agreement must be realized. Ultimately, cross-lingual sharing is driven not by language identity, but by whether the task demands the same underlying grammatical operation.

5.3 Shared heads show agreement-focused and cross-lingually similar routing

Activation patching identifies which heads are causally important for agreement, but not how those heads behave once implicated. We therefore analyze the attention patterns of top-ranked heads to ask two mechanistic questions: whether routing differs across metric and language groups, and whether agreement-related routing is preserved across languages.

Attention profiles clearly separate conjugating and non-conjugating languages. We summarize the attention mass directed from the final prefix position toward the pronoun, the infinitive, and other prompt tokens across three groups: heads identified by the minimal-

pair metric in conjugating languages, and heads identified by the target-logit metric in both conjugating and non-conjugating languages.

As shown in Figure 4 (left), the target-logit metric reveals a stark contrast between language types. In non-conjugating languages, target-logit heads direct less attention to pronoun tokens (0.177) and more to other prompt material (0.762). Conversely, both conjugating groups behave almost identically, placing greater emphasis on the pronoun (minimal-pair: 0.288; target-logit: 0.291) and less on other tokens (minimal-pair: 0.689; target-logit: 0.685).

This functional similarity within conjugating languages provides critical context for the patching results in Section 5.1. Although the target-logit metric surfaces more language-specific head locations than the minimal-pair metric, the implicated heads perform the same broad routing functions. Ultimately, the choice of metric affects how strongly the recovered circuitry is shared across languages, but the fundamental information-routing behavior required for conjugation remains consistent.

Shared heads exhibit similar attention-role profiles across languages. We next ask whether high-impact heads shared across two languages exhibit similar functional profiles in both languages. As shown in Figure 4 (right), cross-lingual role similarity is exceptionally high under both metrics (mean similarity: minimal-pair ≈ 0.928, target-logit ≈ 0.926, based on thousands of directional language-pair observations). This strong cross-lingual functional consistency strengthens the mechanistic interpretation of our patching results. It demonstrates that cross-lingual overlap is not merely a coincidence of layer–head localization, it is consistent with partially shared information-routing roles. Furthermore, even though the two metrics isolate slightly different networks of relevant heads, the specific routing patterns within those networks remain highly consistent across conjugating languages.

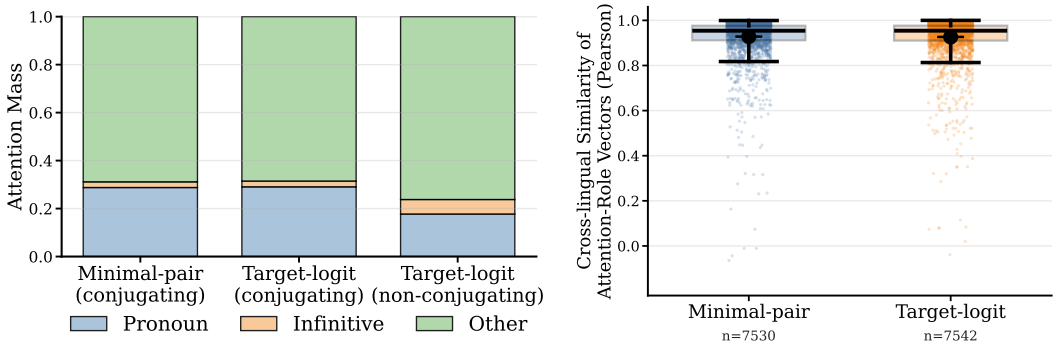


Figure 4: **Shared heads are agreement-focused and functionally similar across languages. Left:** Coarse attention-role composition of top heads across three groups: minimal-pair heads in conjugating languages, target-logit heads in conjugating languages, and target-logit heads in non-conjugating languages. The two conjugating groups show similar coarse routing profiles, whereas target-logit heads in non-conjugating languages allocate less attention to pronouns and more to other prompt material. **Right:** Cross-lingual similarity of attention-role vectors for overlapping top-20 heads, shown separately by metric. High similarity under both metrics indicates that implicated heads tend to preserve agreement-related routing roles across languages.

6 Discussion

We investigated whether multilingual language models implement present-tense subject–verb agreement through shared or language-specific internal circuitry. Across 29 languages and five model families, we found evidence for a middle-ground view: multilingual models often reuse partially shared circuitry for agreement, but this sharing is neither uniform nor fully language-agnostic. Languages with overt agreement morphology exhibit substantially more similar head-level signatures than comparisons involving non-conjugating languages, and many implicated heads also show similar attention behavior across languages. Together,

these results suggest that cross-lingual overlap in agreement circuitry reflects not only shared localization, but also partially shared functional roles.

These findings extend prior demonstrations that cross-lingual circuit overlap can occur by identifying conditions under which that overlap strengthens or weakens: sharing is shaped less by language identity alone than by whether the same grammatical computation must be overtly realized. This is clearest in the contrast between our two recovery metrics: sharing is strongest when the analysis isolates recovery of the inflectional contrast itself, and weaker when it targets broader context-appropriate form recovery. English provides especially strong evidence for this view. It does not behave like a strongly conjugating language overall, but its third-person singular cases become more similar to conjugating languages precisely in the contexts where overt agreement must be computed. This suggests that multilingual models reuse shared circuitry most readily when languages instantiate the same agreement-relevant operation, rather than simply because they belong to the same broad language class.

More broadly, our findings suggest that shared multilingual computation is shaped not only by semantics, but also by whether languages realize the same structural feature in comparable ways. This has implications for multilingual interpretability, as well as for when causal interventions or alignment methods developed in one language should be expected to transfer to others.

These findings should be interpreted in light of several limitations. Our analysis characterizes the attention-based routing component of agreement rather than the full computational graph: although attention heads enable the inter-token transfer of subject information, position-wise MLPs may play important downstream roles in reading, transforming, or amplifying this signal. Moreover, because we use structured conjugation prompts rather than naturalistic sentences, the identified attention-head signatures may not fully generalize to agreement processing in ordinary language contexts. We also focus on one morphosyntactic phenomenon, on attention-head signatures rather than full circuit reconstruction, and on filtered cases in which the model successfully produces the target form. Extending this analysis to additional grammatical phenomena, model components, and training settings is an important direction for future work. More generally, we hope this work helps shift multilingual mechanistic interpretability from documenting whether internal structure is shared to explaining when grammatical computations recruit shared rather than language-specific mechanisms.

Acknowledgments

Part of this work was carried out during the JSALT 2025 workshop as part of the Play-your-Part team. The workshop was supported by discretionary funds from Johns Hopkins University and by the Ministry of Education, Youth and Sports of the Czech Republic through the OP JAK project "Linguistics, Artificial Intelligence and Language and Speech Technologies: From Research to Applications" (project ID CZ.02.01.01/00/23 020/0008518).

This work received funding from the French National Research Agency through grants ANR-16-CONV-0002 (Institute for Language, Communication and the Brain; ILCB) and ANR-20-CE23-0012-01 MIM; the European Union's Horizon 2020 research and innovation programme under Marie Skłodowska-Curie grant agreement No. 101007666; MCIN/AEI/10.13039/501100011033 under grant PID2024-155948OB-C53; and the Government of Aragon under grant T36 23R. This work was also enabled in part by a gift from the Chan Zuckerberg Initiative Foundation to establish the Kempner Institute for the Study of Natural and Artificial Intelligence.

References

Khuyagbaatar Batsuren, Omer Goldman, Salam Khalifa, Nizar Habash, Witold Kieraś, Gábor Bella, Brian Leonard, Garrett Nicolai, Kyle Gorman, Yustinus Ghanggo Ate, et al. UniMorph 4.0: Universal Morphology. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara,

Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, and Stelios Piperidis (eds.), *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pp. 840–855, Marseille, France, June 2022. European Language Resources Association. URL https://aclanthology.org/2022.lrec-1.89/.

BigScience Workshop, Teven Le Scao, et al. BLOOM: A 176B-parameter open-access multilingual language model, 2022. URL https://arxiv.org/abs/2211.05100.

Jannik Brinkmann, Chris Wendler, Christian Bartelt, and Aaron Mueller. Large language models share representations of latent grammatical concepts across typologically diverse languages. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *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)*, pp. 6131–6150, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.312. URL https://aclanthology.org/2025.naacl-long.312/.

Boyi Deng, Yu Wan, Baosong Yang, Yidan Zhang, and Fuli Feng. Unveiling language-specific features in large language models via sparse autoencoders. 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)*, pp. 4563–4608, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.229. URL https://aclanthology.org/2025.acl-long.229/.

Yaling Deng, Jiwen Li, Minglu Niu, Ye Wang, Wenlong Fu, Yanzhu Gong, Shuo Ding, Wenyi Li, Wei He, and Lihong Cao. A chinese verb semantic feature dataset (CVFD). *Behavior Research Methods*, 56(1):342–361, 2024. doi: 10.3758/s13428-022-02047-4. URL https://doi.org/10.3758/s13428-022-02047-4.

Clément Dumas, Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. Separating tongue from thought: Activation patching reveals language-agnostic concept representations in transformers. 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)*, pp. 31822–31841, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1536. URL https://aclanthology.org/2025.acl-long.1536/.

Javier Ferrando and Marta R. Costa-jussà. On the similarity of circuits across languages: a case study on the subject-verb agreement task. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 10115–10125, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.591. URL https://aclanthology.org/2024.findings-emnlp.591/.

Gemma Team. Gemma 2: Improving open language models at a practical size, 2024. URL https://arxiv.org/abs/2408.00118.

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

Albert Q. Jiang et al. Mistral 7b, 2023. URL https://arxiv.org/abs/2310.06825.

Yi Jing, Zijun Yao, Hongzhu Guo, Lingxu Ran, Xiaozhi Wang, Lei Hou, and Juanzi Li. LinguaLens: Towards interpreting linguistic mechanisms of large language models via sparse auto-encoder. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 28232–28251, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1433. URL https://aclanthology.org/2025.emnlp-main.1433/.

Jiwei Li, Xinlei Chen, Eduard Hovy, and Dan Jurafsky. Visualizing and understanding neural models in NLP. In Kevin Knight, Ani Nenkova, and Owen Rambow (eds.), *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 681–691, San Diego, California, June 2016. Association for Computational Linguistics. doi: 10.18653/v1/N16-1082. URL https://aclanthology.org/N16-1082/.

Vladislav Mikhailov, Oleg Serikov, and Ekaterina Artemova. Morph call: Probing morphosyntactic content of multilingual transformers. In Ekaterina Vylomova, Elizabeth Salesky, Sabrina Mielke, Gabriella Lapesa, Ritesh Kumar, Harald Hammarström, Ivan Vulić, Anna Korhonen, Roi Reichart, Edoardo Maria Ponti, and Ryan Cotterell (eds.), *Proceedings of the Third Workshop on Computational Typology and Multilingual NLP*, pp. 97–121, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.sigtyp-1.10. URL https://aclanthology.org/2021.sigtyp-1.10/.

Aaron Mueller, Yu Xia, and Tal Linzen. Causal analysis of syntactic agreement neurons in multilingual language models. In Antske Fokkens and Vivek Srikumar (eds.), *Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL)*, pp. 95–109, Abu Dhabi, United Arab Emirates (Hybrid), December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.conll-1.8. URL https://aclanthology.org/2022.conll-1.8/.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. Crosslingual generalization through multitask finetuning. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15991–16111, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.891. URL https://aclanthology.org/2023.acl-long.891/.

Karolina Stanczak, Edoardo Ponti, Lucas Torroba Hennigen, Ryan Cotterell, and Isabelle Augenstein. Same neurons, different languages: Probing morphosyntax in multilingual pre-trained models. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz (eds.), *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1589–1598, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.114. URL https://aclanthology.org/2022.naacl-main.114/.

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)*, pp. 5701–5715, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.309. URL https://aclanthology.org/2024.acl-long.309/.

Eric Todd, Millicent 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*, 2024. URL https://openreview.net/forum?id=AwyxtyMwaG.

Hugo Touvron et al. Llama 2: Open foundation and fine-tuned chat models, 2023. URL https://arxiv.org/abs/2307.09288.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 12388–12401. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/92650b2e92217715fe312e6fa7b90d82-Paper.pdf.

Xinpeng Wang, Mingyang Wang, Yihong Liu, Hinrich Schuetze, and Barbara Plank. Refusal direction is universal across safety-aligned languages. In *Advances in Neural Information Processing Systems*, volume 38, pp. 32380–32423. Curran Associates, Inc., 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/2e94772e3c079d83f79c311d07456111-Abstract-Conference.html.

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)*, pp. 15366–15394, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.820. URL https://aclanthology.org/2024.acl-long.820/.

An Yang et al. Qwen2 technical report, 2024. URL https://arxiv.org/abs/2407.10671.

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*, 2024. URL https://openreview.net/forum?id=Hf17y6u9BC.

Ruochen Zhang, Qinan Yu, Matianyu Zang, Carsten Eickhoff, and Ellie Pavlick. The same but different: Structural similarities and differences in multilingual language modeling. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=NCrFA7dq8T.

A Task and Experimental Details

A.1 Conjugation task and agreement patterns

We study present-tense subject–verb agreement through controlled contrasts in grammatical person and number. Our goal is to isolate when a change in the subject requires a change in the surface form of the verb, and to compare how this computation is implemented across languages. Figure 5 shows representative examples from three typologically distinct cases: English, which marks agreement only in the third-person singular; Spanish, which exhibits rich person/number inflection; and Swedish, which does not mark person in the tested present-tense contexts.

For reference, Figure 6 lists the full language inventory drawn from our datasets, and Table 3 groups the languages retained for analysis by whether the tested contrasts involve overt agreement marking. This organization is important for the main comparisons in the paper, since our cross-lingual analyses distinguish languages with overt agreement from languages where the same contrasts are not morphologically realized.

Grammatical Person	**Pronoun**	**English - "speak"**	**Spanish - "hablar"**	**Swedish - "talar"**
First Person Singular	I	speak	habl**o**	talar
Second Person Singular	you	speak	habl**as**	talar
Third Person Singular	he/she	speak**s**	habl**a**	talar
First Person Plural	we	speak	habl**amos**	talar
Second Person Plural	you all	speak	habl**áis**	talar
Third Person Plural	they	speak	habl**an**	talar

Figure 5: Contrastive examples of verb conjugation across three typologically distinct languages. English only inflects verbs in the third-person singular. Spanish exhibits rich person/number inflection. Swedish does not inflect for person in the tested present-tense contexts.

A.2 Models used in the main experiments

We evaluate the main analyses on a diverse set of open-source multilingual model families. This set gives us variation in architecture, training corpus, and scale while keeping the experimental pipeline fixed across models.

Table 1: Models used in the main experiments.

Family	**Hugging Face Name**	**Citation**
BLOOM	bigscience/bloom-7b1	(BigScience Workshop et al., 2022)
Gemma	google/gemma-2-9b	(Gemma Team, 2024)
Llama	Llama-2-7b-hf	(Touvron et al., 2023)
	meta-llama/Llama-3.2-3B	(Grattafiori et al., 2024)
Mistral	mistralai/Mistral-7B-v0.1	(Jiang et al., 2023)
Qwen	Qwen/Qwen2-7B	(Yang et al., 2024)

A.3 Prompt templates and pronoun inventories

For each language, we use a fixed zero-shot prompt template designed to elicit the target present-tense form. The subject pronoun is always stated explicitly, ensuring that the

agreement cue is overtly available in every condition. This helps make the cross-lingual comparisons more interpretable: when models differ, the differences are less likely to be driven by whether the subject cue was present in the prompt at all.

Figure 6 displays the full language-specific prompt templates, and the following figure lists the pronoun inventories used for the tested person and number contrasts. Together, these materials define the multilingual prompt space from which the clean/corrupted minimal pairs are constructed.

Language	**Prompt template**
Albanian	Zgjedhimi i foljes {verb} në kohën e tashme: {pronoun} {conj}
Bulgarian	Спрежение на глагола {verb} в сегашно време: {pronoun} {conj}
Catalan	Conjugació del verb {verb} en present: {pronoun} {conj}
Chinese	动词{verb}（现在时）： {pronoun} {conj}
Czech	Časování slovesa {verb} v přítomném čase: {pronoun} {conj}
Danish	Bøjning af verbet {verb} i nutid: {pronoun} {conj}
English	Conjugation of the verb {verb} in present tense: {pronoun} {conj}
Estonian	Tegusõna {verb} pööramine olevikus: {pronoun} {conj}
Finnish	Verbin {verb} taivutus preesensissä: {pronoun} {conj}
French	Conjugaison du verbe {verb} au présent : {pronoun} {conj}
German	Konjugation des Verbs {verb} im Präsens: {pronoun} {conj}
Hindi	वर्तमान काल में क्रिया {verb}का रूप: {pronoun} {conj}
Hungarian	A(z) {verb} ige ragozása jelen időben: {pronoun} {conj}
Icelandic	Beyging sagnarinnar {verb} í nútíð: {pronoun} {conj}
Irish	Réimniú an bhriathair {verb} san aimsir láithreach: {pronoun} {conj}
Italian	Coniugazione del verbo {verb} al presente: {pronoun} {conj}
Latin	Coniugatio verbi {verb} in praesenti: {pronoun} {conj}
Latvian	Darbības vārda {verb} locīšana tagadnē: {pronoun} {conj}
Lithuanian	Veiksmažodžio {verb} asmenavimas esamuoju laiku: {pronoun} {conj}
Modern Greek	Κλίση του ρήματος {verb} στον ενεστώτα: {pronoun} {conj}
Norwegian Bokmål	Bøyning av verbet {verb} i presens: {pronoun} {conj}
Persian	فعل صرف {verb}در زمان: حال {pronoun} {conj}
Polish	Odmiana czasownika {verb} w czasie teraźniejszym: {pronoun} {conj}
Portuguese	Conjugação do verbo {verb} no presente: {pronoun} {conj}
Romanian	Conjugarea verbului {verb} la prezent: {pronoun} {conj}
Russian	Спряжение глагола {verb} в настоящем времени: {pronoun} {conj}
Serbo-Croatian	Konjugacija glagola {verb} u prezentu: {pronoun} {conj}
Slovenian	Spreganje glagola {verb} v sedanjiku: {pronoun} {conj}
Spanish	Conjugación del verbo {verb} en presente: {pronoun} {conj}
Swedish	Böjning av verbet {verb} i presens: {pronoun} {conj}
Turkish	{verb} fiilinin şimdiki zamandaki çekimi: {pronoun} {conj}
Ukrainian	Відмінювання дієслова {verb} у теперішньому часі: {pronoun} {conj}
Urdu	فعل {verb}کی میں حال زمانہ: تصریف {pronoun} {conj}
Welsh	Cydgysylltiad y ferf {verb} yn yr amser presennol: {pronoun} {conj}

Figure 6: Language-specific prompt templates for all tested languages.

Language	1sg	2sg	3sg	1pl	2pl	3pl
Albanian	Unë	Ti	Ai	Ne	Ju	Ata
Bulgarian	Аз	Ти	Той	Ние	Вие	Те
Catalan	Jo	Tu	Ell	Nosaltres	Vosaltres	Ells
Chinese	我	你	他	我们	你们	他们
Czech	Já	Ty	On	My	Vy	Oni
Danish	Jeg	Du	Han	Vi	I	De
English	I	You	He	We	You	They
Estonian	Mina	Sina	Tema	Meie	Teie	Nemad
Finnish	Minä	Sinä	Hän	Me	Te	He
French	Je	Tu	Il	Nous	Vous	Ils
German	Ich	Du	Er	Wir	Ihr	Sie
Hindi	मैं	तुम	वह	हम	तुम लोग	वे
Hungarian	Én	Te	Ő	Mi	Ti	Ők
Icelandic	Ég	Þú	Hann	Við	Þið	Þeir
Irish	Mé	Tú	Sé	Muid	Sibh	Siad
Italian	Io	Tu	Lui	Noi	Voi	Loro
Latin	Ego	Tu	Is	Nos	Vos	Ii
Latvian	Es	Tu	Viņš	Mēs	Jūs	Viņi
Lithuanian	Aš	Tu	Jis	Mes	Jūs	Jie
Modern Greek	Εγώ	Εσύ	Αυτός	Εμείς	Εσείς	Αυτοί
Norwegian Bokmål	Jeg	Du	Han	Vi	Dere	De
Persian	من	تو	او	ما	شما	آنها
Polish	Ja	Ty	On	My	Wy	Oni
Portuguese	Eu	Tu	Ele	Nós	Vós	Eles
Romanian	Eu	Tu	El	Noi	Voi	Ei
Russian	Я	Ты	Он	Мы	Вы	Они
Serbo-Croatian	Ja	Ti	On	Mi	Vi	Oni
Slovenian	Jaz	Ti	On	Mi	Vi	Oni
Spanish	Yo	Tú	Él	Nosotros	Vosotros	Ellos
Swedish	Jag	Du	Han	Vi	Ni	De
Turkish	Ben	Sen	O	Biz	Siz	Onlar
Ukrainian	Я	Ти	Він	Ми	Ви	Вони
Urdu	میں	تم	وہ	ہم	لوگ تم	لوگ وہ
Welsh	Fi	Ti	Fe	Ni	Chi	Nhw

Figure 7: Subject pronouns used in each tested language.

A.4 Tokenization Filtering

To isolate the agreement contrast as tightly as possible, we retain only cases in which the clean and corrupted conjugated forms differ in the final token and otherwise remain aligned. This model-specific filtering step is crucial for the patching analysis: it ensures that the manipulated contrast is localized to the token being predicted rather than spread across multiple positions in the sequence.

Figure 8 illustrates representative retained and discarded cases. Retained examples preserve the shared prefix and differ only where the agreement contrast should matter most, whereas discarded examples introduce broader tokenization differences that would confound causal interpretation.

```
----------------------------------------
Infinitive: comer
  Yo: como -> ['como']
  Tú: comes -> ['comes']
----------------------------------------
----------------------------------------
Infinitive: meter
  Yo: meto -> ['met', 'o']
  Tú: metes -> ['met', 'es']
----------------------------------------
----------------------------------------
Infinitive: esconder
  Yo: escondo -> ['es', 'con', 'do']
  Tú: escondes -> ['es', 'con', 'des']
----------------------------------------

----------------------------------------
Infinitive: conocer
  Yo: conozco -> ['con', 'oz', 'co']
  Tú: conoces -> ['con', 'oces']
----------------------------------------
```

Figure 8: Examples retained and discarded by the tokenization filter.

A.5 Languages Tested

A.5.1 Languages In Dataset

We begin from a broad multilingual inventory assembled from UniMorph and the Chinese Verb Semantic Feature Dataset. Not every language survives the later filtering steps in every model, but this initial inventory gives us wide typological coverage and includes both overtly conjugating and non-conjugating languages.

Languages
Albanian, Bulgarian, Catalan, Chinese, Czech, Danish, English, Estonian, Finnish, French, German, Hindi, Hungarian, Icelandic, Irish, Italian, Latin, Latvian, Lithuanian, Modern Greek, Norwegian Bokmål, Persian, Polish, Portuguese, Romanian, Russian, Serbo-Croatian, Slovenian, Spanish, Swedish, Turkish, Ukrainian, Urdu, Welsh

Table 2: Languages tested across all models (pulled from dataset). Not all of these languages were used in analysis if they did not pass the filtering.

A.5.2 Languages Used In Analysis

After tokenization and behavioral filtering, the analysis set partitions naturally into three groups: languages with overt agreement in the tested contrasts, languages without overt agreement in those contrasts, and English as a middle case. This grouping is the basis for the cross-group comparisons in the main text and appendix.

Group	Languages
Conjugating	Albanian, Bulgarian, Catalan, Czech, Estonian, Finnish, French, German, Hindi, Hungarian, Icelandic, Italian, Latin, Latvian, Lithuanian, Modern Greek, Polish, Portuguese, Romanian, Russian, Serbo-Croatian, Spanish, Turkish, Urdu
Nonconjugating	Chinese, Danish, Norwegian Bokmål, Swedish
Middle	English

Table 3: Languages that survive the filtering and appear in at least one patch result (in either the target-logit or minimal-pair recovery) in at least one model, grouped by agreement type. These are the ones used in analysis.

A.5.3 Language-model coverage

Official language support varies across families, but we retain a language–model combination whenever at least 50 prompts in that language remain after tokenization and behavioral filtering. Figure 9 shows that coverage is broad but uneven across models. Gemma, Mistral, and Qwen retain especially wide coverage, whereas BLOOM and the Llama models exhibit more gaps after filtering. Even so, every model contributes a substantial multilingual set, allowing the main cross-lingual comparisons to be matched within model family.

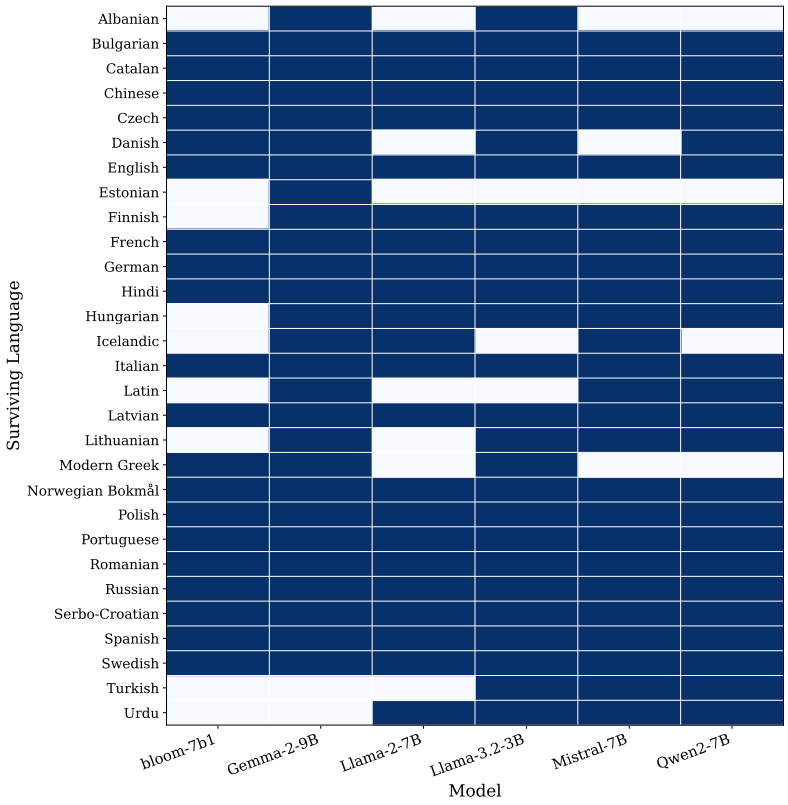


Figure 9: Languages surviving for tested models.

B Supplementary Results for Main Claims

B.1 Model-specific and individual language-level cross-lingual similarity among conjugating languages.

Figure 10 expands the main-text result that conjugating languages exhibit positive cross-lingual similarity under the minimal-pair metric. The central pattern is broad, not sparse: within each family, the similarity matrices are dominated by positive values rather than by a few isolated high-similarity pairs. In other words, the effect is not carried by only one language cluster or one especially favorable comparison.

The strength of the pattern does vary by family. Gemma, Mistral, and Qwen show especially dense blocks of positive similarity, while BLOOM and Llama display somewhat more heterogeneity and more missing cells. Importantly, the white cells mark missing comparisons due to coverage constraints rather than evidence against cross-lingual sharing. Overall, these matrices reinforce the claim that shared agreement circuitry among conjugating languages is a broad multilingual phenomenon rather than a narrow family-specific artifact.

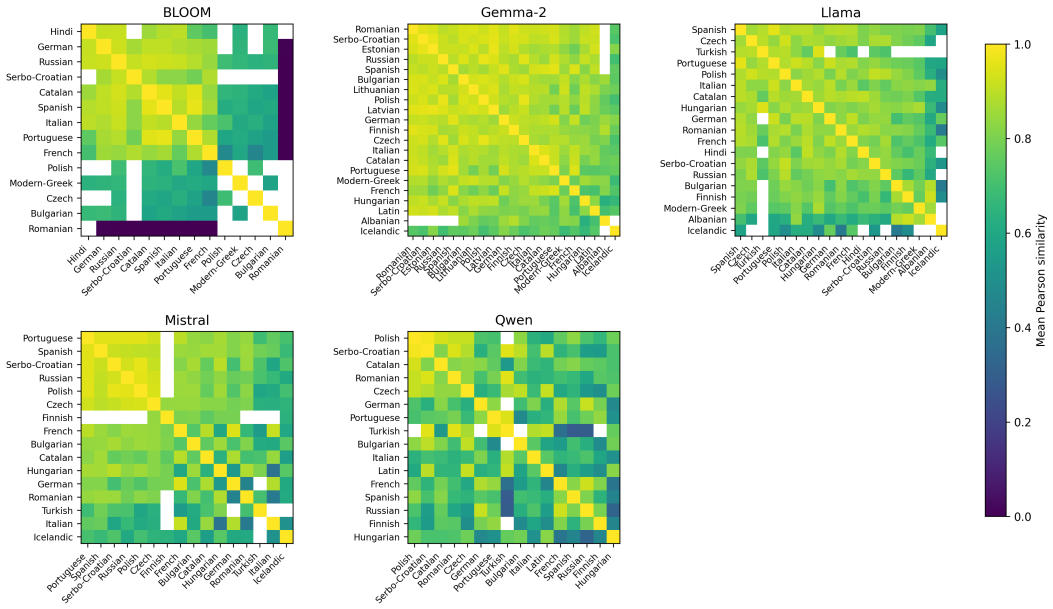


Figure 10: Family-specific language-similarity matrices for conjugating languages under the minimal-pair recovery metric. Each panel shows mean pairwise Pearson similarity between flattened absolute heatmaps, computed at the model-pair level and then averaged within family. White cells denote missing comparisons.

B.2 Similarity by Conjugation Contrast and Patching Direction

To test whether the cross-lingual sharing observed in the main analysis is driven by particular agreement contrasts or patching directions, we disaggregate minimal-pair similarity by conjugation pair and direction. Table 4 reports the mean pairwise Pearson similarity between conjugating languages, averaged across model checkpoints. Cross-lingual similarity remains positive in every condition, indicating that the main pattern is not restricted to a particular patching direction, although the magnitude of similarity varies across contrasts.

Conjugation	Direction	Mean	SD	n
1sg–2sg	1sg → 2sg	0.6889	0.2950	1459
1sg–2sg	2sg → 1sg	0.6331	0.3269	1483
1sg–3sg	1sg → 3sg	0.6045	0.3017	1476
1sg–3sg	3sg → 1sg	0.6388	0.2887	1508
1pl–3pl	1pl → 3pl	0.5177	0.3467	1290
1pl–3pl	3pl → 1pl	0.5040	0.3312	1320
3sg–3pl	3pl → 3sg	0.5628	0.3247	1558
3sg–3pl	3sg → 3pl	0.5362	0.3039	1528
All	All directions	0.5876	0.3204	11622

Table 4: Cross-lingual activation-patching similarity among conjugating languages, disaggregated by conjugation contrast and patching direction, where n denotes the number of matched language-pair observations contributing to each condition. Values report mean pairwise Pearson similarity between absolute minimal-pair recovery heatmaps, averaged across model checkpoints. The positive similarity observed in the aggregate analysis persists across all individual directions, although its magnitude varies across contrasts.

B.3 Model-specific cross-group comparisons

Figure 11 breaks out the main cross-group comparison by model family. The same ordering appears in every family: conjugating–conjugating comparisons are highest, conjugating–non-conjugating comparisons are lower, and non-conjugating–non-conjugating comparisons remain lowest overall. This consistency matters because it shows that the grammatical separation highlighted in the main text is not being driven by a single architecture. Taken together, these panels show that the cross-group separation is stable across model families, even though the magnitude of the gap varies.

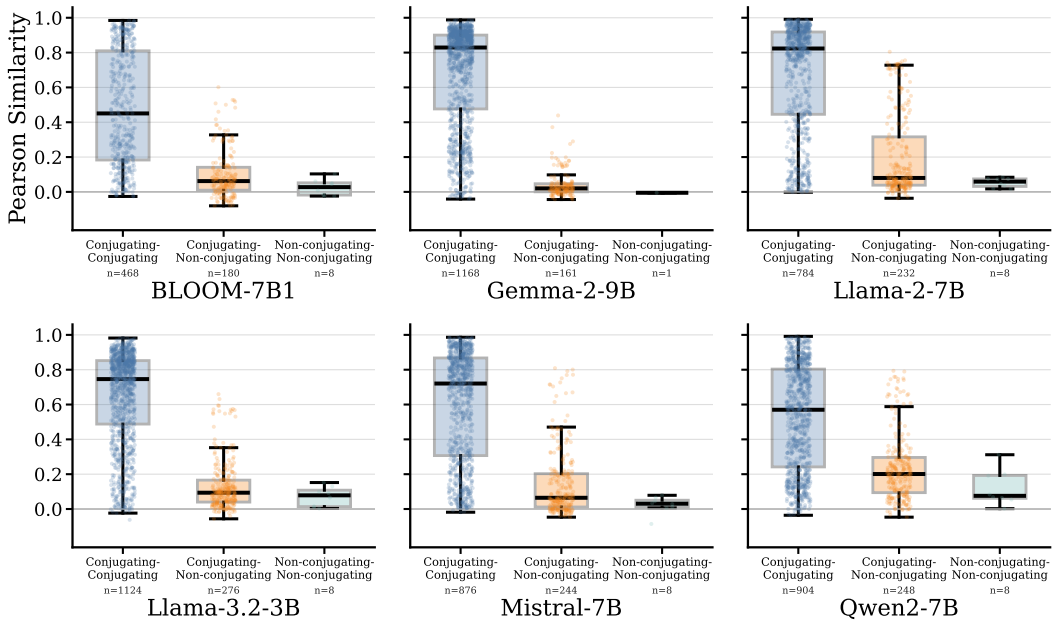


Figure 11: Model-specific cross-group similarity distributions. Within each panel, conjugating–conjugating comparisons are consistently more similar than conjugating–non-conjugating comparisons, although the size of the gap varies by model family.

B.4 Additional analyses of English's mixed agreement profile

Figure 12 provides a model-by-model view of the English effect from the main text. The result is directionally consistent across all six models: English is always more similar to conjugating-language circuitry in the 3sg-inflecting conditions than in the non-inflecting conditions. What changes across models is the size of the shift, not its sign.

The largest shifts appear in Gemma-2-9B and BLOOM-7B1, with Llama-3.2-3B and Qwen2-7B also showing clear positive differences. Mistral-7B and Llama-2-7B exhibit smaller but still positive shifts. This model-level consistency strengthens the interpretation that English behaves as a true bridge case: it aligns more closely with conjugating languages precisely in the contexts where English itself overtly realizes agreement.

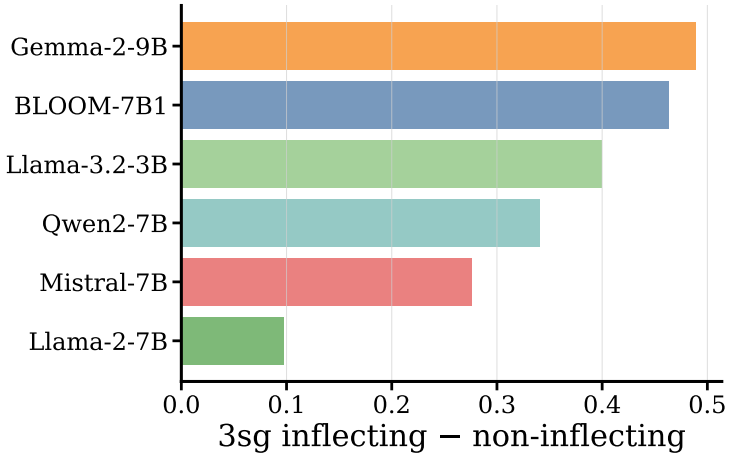


Figure 12: English shift toward conjugating-language circuitry, by model.

C Robustness to Similarity Metric

Figure 13 compares the cross-lingual similarity result across several alternative similarity metrics and head-selection schemes. The broad conclusion is that the main pattern is robust, but some metrics capture it more cleanly than others. Pearson and cosine similarity reproduce the effect most stably across both all-head and top-k variants, while Spearman is visibly noisier and less discriminating. Jaccard overlap on the top-20 heads captures some shared structure, but compresses the distribution and loses graded information relative to correlation-based measures.

For this reason, the main text focuses on Pearson correlation over flattened absolute heatmaps. It preserves the continuous structure of the patching signature and yields the clearest separation between stronger and weaker cross-lingual overlap. The robustness figure therefore supports the substantive claim while also motivating the specific similarity metric used in the paper.

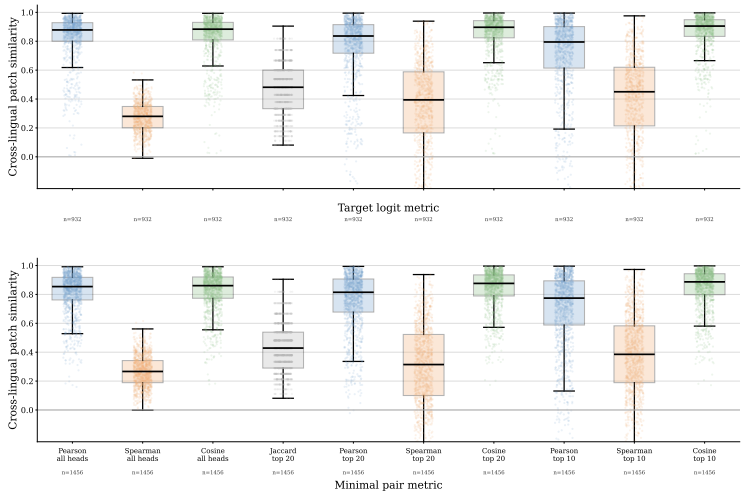


Figure 13: Cross-lingual similarity of conjugating languages under alternative similarity metrics and head-selection schemes. Pearson and cosine recover the main pattern most consistently, whereas Spearman is visibly noisier.

D Auxiliary Analyses

For the auxiliary analyses, we work with a smaller language subset (Table 5) chosen to preserve typological diversity while making the expanded model comparisons computationally tractable. This subset includes richly conjugating languages, English as a mixed case, and non-conjugating comparison languages.

Languages
Spanish, Portuguese, Italian, French, Catalan, Czech, Finnish, Russian, Hungarian, Polish, Serbo-Croatian, German, English, Swedish, Chinese, Mongolian

Table 5: Languages included in the auxiliary experiments.

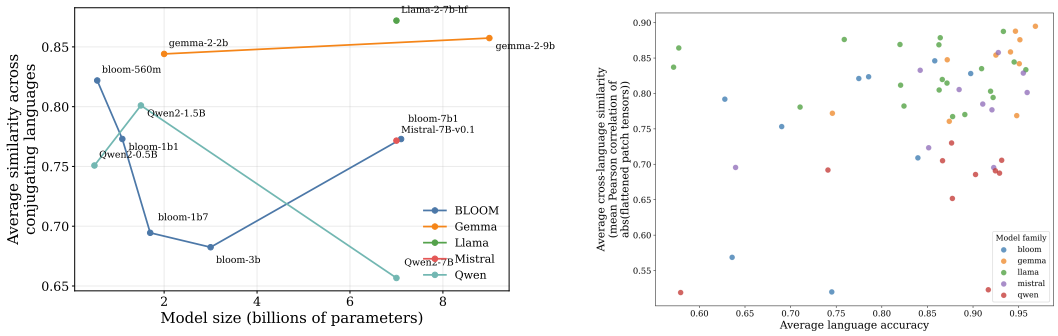
D.1 Model size and language accuracy

These analyses test whether the cross-lingual similarity pattern could be reduced to either model scale or language-specific task performance. Figure 14 suggests that neither account is sufficient on its own.

The model-scale analysis does not show a single monotonic relationship between parameter count and cross-lingual similarity. Within some families, similarity rises with scale, but in others it flattens or changes only modestly. Likewise, the accuracy analysis shows that better language-specific task performance does not mechanically imply stronger cross-lingual circuit similarity. Languages and models with similar accuracy can still differ noticeably in cross-language overlap. These results support the view that the main similarity pattern reflects something more specific than generic model quality or scale.

Table 6: Models used in the model-scale and accuracy robustness analyses.

Family	Hugging Face Name	Citation
BLOOM	bigscience/bloom-560m	(BigScience Workshop et al., 2022)
BLOOM	bigscience/bloom-1b1	(BigScience Workshop et al., 2022)
BLOOM	bigscience/bloom-1b7	(BigScience Workshop et al., 2022)
BLOOM	bigscience/bloom-3b	(BigScience Workshop et al., 2022)
BLOOM	bigscience/bloom-7b1	(BigScience Workshop et al., 2022)
Gemma	google/gemma-2-9b	(Gemma Team, 2024)
Gemma	google/gemma-2-2b	(Gemma Team, 2024)
Llama	Llama-2-7b-hf	(Touvron et al., 2023)
Mistral	mistralai/Mistral-7B-v0.1	(Jiang et al., 2023)
Qwen	Qwen/Qwen2-0.5B	(Yang et al., 2024)
Qwen	Qwen/Qwen2-1.5B	(Yang et al., 2024)
Qwen	Qwen/Qwen2-7B	(Yang et al., 2024)



(a) Cross-lingual similarity as a function of model scale.

(b) Cross-lingual similarity versus language-specific task accuracy.

Figure 14: Robustness checks for model scale and task accuracy.

D.2 Instruction tuning

We also test whether the recovered similarity structure changes coherently after instruction tuning. Figure 15 shows that the answer is largely no: instruction tuning does not introduce a consistent upward or downward shift in cross-lingual similarity across families.

Any changes that do appear are modest and mixed in direction. This suggests that the circuit-level cross-lingual structure relevant to agreement is already present in the base models and is not substantially reorganized by instruction tuning. In other words, the sharing pattern documented in the main text appears to be relatively stable under this form of post-training.

Base	Instruction-tuned	Citation
bloom-1b7	bloomz-1b7	(BigScience Workshop et al., 2022; Muennighoff et al., 2023)
gemma-2-2b	gemma-2-2b-it	(Gemma Team, 2024)
Llama-2-7b-hf	Llama-2-7b-chat-hf	(Touvron et al., 2023)
Mistral-7B-v0.1	Mistral-7B-Instruct-v0.1	(Jiang et al., 2023)
Qwen2-1.5B	Qwen2-1.5B-Instruct	(Yang et al., 2024)

Table 7: Base and instruction-tuned models used in the instruction-tuning comparison.

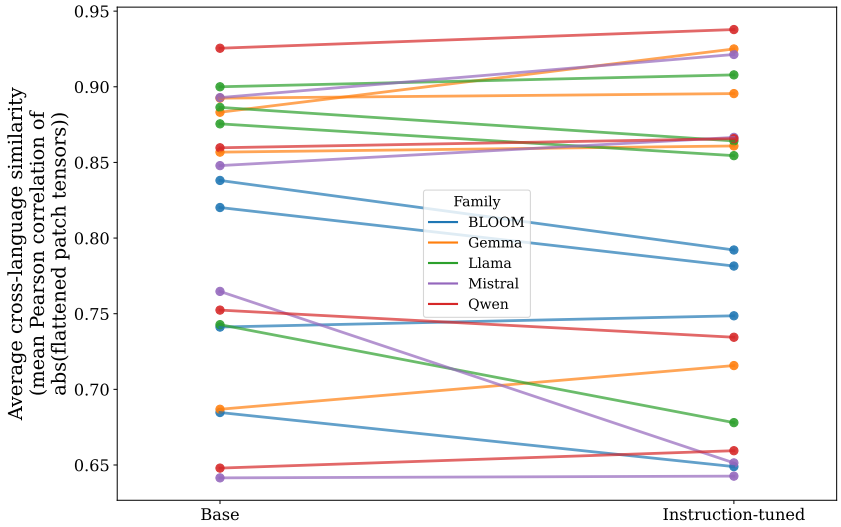


Figure 15: Instruction tuning leaves the overall similarity structure largely unchanged.

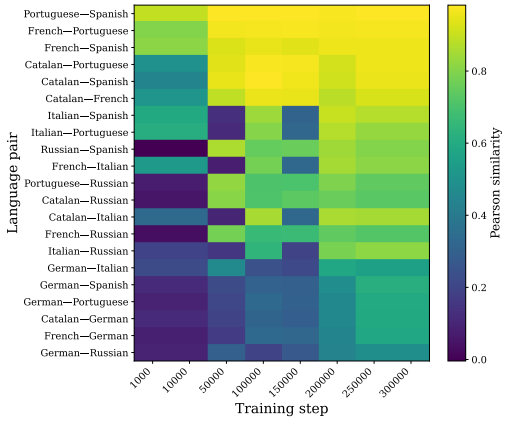
D.3 Analyses of pre-training progression

As a training-dynamics case study, we examine intermediate BLOOM checkpoints to ask whether cross-lingual similarity in conjugation circuitry emerges gradually over pretraining. Figure 16 shows a clear overall trend: across both recovery metrics, similarity tends to increase over training, although the minimal-pair signal is noisier and more contrast-specific than the target-logit signal.

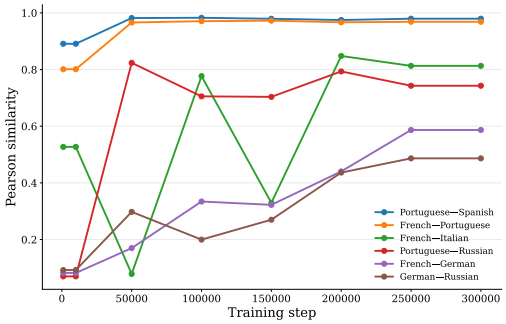
The language-pair heatmap and representative trajectories show that individual pairs do not all move identically. Some pairs rise sharply early and then plateau, while others increase more gradually. But at the aggregate level, the trajectory is upward, indicating that shared cross-lingual agreement circuitry is not fully present from the start and instead becomes more pronounced over pretraining. The comparison between the two metrics further mirrors the main-text story: the target-logit metric produces a smoother developmental trajectory, whereas the minimal-pair metric captures a more selective and variable inflection-specific signal.

Base model	`bigscience/bloom-1b7`
Checkpoint repo	`bigscience/bloom-1b7-intermediate`
Steps	1k, 10k, 50k, 100k, 150k, 200k, 250k, 300k

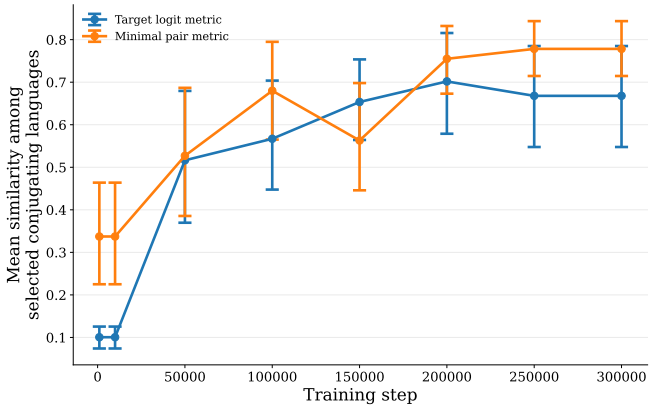
Table 8: Base model and intermediate checkpoints used in the BLOOM progression analysis (BigScience Workshop et al., 2022).



(a) Similarity by language pair and checkpoint in the minimal-pair metric.



(b) Representative language-pair trajectories in the minimal-pair metric.



(c) Comparison of target-logit and minimal-pair similarity trajectories.

Figure 16: Supplementary views of the BLOOM checkpoint analysis. Across both metrics, cross-lingual similarity tends to increase over training, although the minimal-pair metric is noisier and more contrast-specific.