WHAT MULTILINGUAL MODELS SHARE: THE FUNCTIONAL AND NEURAL SIGNIFICANCE OF REPRESENTATIONAL CONVERGENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Multilingual models transfer knowledge across languages, offering a way to study shared principles of language organization. Interpreting representational convergence requires an account of how shared structure relates to computation and human processing. We compare 60 languages across nine models from five families and examine the functional and neural significance of shared organization using grammatical interventions and Chinese intracranial recordings. Relations among aligned meanings recur more strongly across model families than activation organization or sequential response strength, while local reconstruction depends on the source language. Edits to grammatical-number information selectively weaken English agreement for forms excluded from estimation in every model tested, and later edits offset the loss in most models; confined to the units that carry the recurrent relations, the edit retains 3–41% of its effect, within or below the range of matched random units, so grammatical influence is distributed beyond those units. In Qwen2.5-7B, English- and Hebrew-derived edits produce similar rankings of affected Hebrew sentences, while strength and surrounding number information shape effect magnitude. Chinese intracranial recordings from 21 participants distinguish overall neural resemblance from the additional correspondence to the same sentences, and the shared and model-specific parts of a representation differ in that correspondence by less than one tenth of the neural reliability reference. These findings clarify the functional and neural significance of shared linguistic knowledge, informing theories of multilingual intelligence that accommodate common organization and computational diversity. Code, results, and experimental settings are available at `https://anonymous.4open.science/r/Multilingual-69FC/`.

1 INTRODUCTION

Multilingual models apply knowledge learned in one language to tasks in another (Pires et al., 2019; Conneau et al., 2020), and their representations encode linguistic structure and ancestry (Malaviya et al., 2017; Rama et al., 2020). Representational convergence (Huh et al., 2024) offers a starting point for studying shared principles of language organization. We test how that organization relates to computation and to human language processing, comparing 60 languages across nine models from five families and linking recurring language relations with grammatical interventions and Chinese intracranial responses.

Relating shared organization to computation requires evidence of its role in behavior. Grammatical number provides a concrete case: after *the key near the cabinets*, the verb must agree with the singular head despite the intervening plural noun. Probing asks what can be recovered from a representation (Hewitt & Liang, 2019; Belinkov, 2022); intervention tests how changing it affects behavior (Elazar et al., 2021; Vig et al., 2020; Geiger et al., 2021; 2024); agreement interventions overlap across languages, varying with architecture and construction (Mueller et al., 2022). We compare number edits confined to units selected for language relations with edits confined to random units, on held-out forms of number marking. English-to-Hebrew transfer in Qwen2.5-7B examines how these effects vary with edit strength and surrounding number information.

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

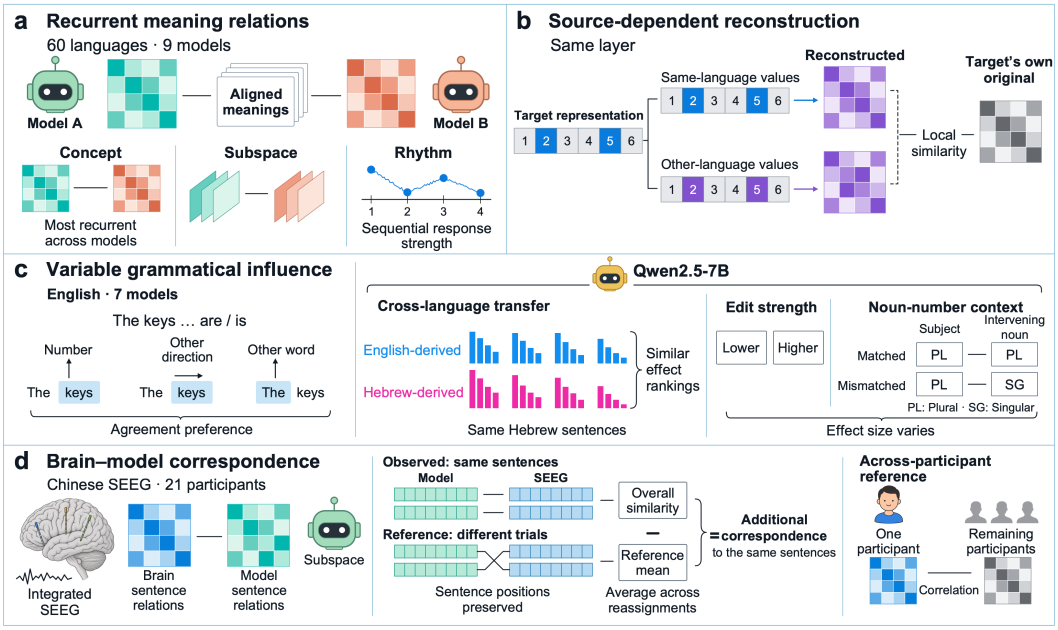

Figure 1: **Recurrent meaning relations and variable grammatical influence.** Panels and electrodes are schematic. **a**, Concept relations recur most strongly. **b**, Source-dependent same-layer reconstruction. **c**, English edits test held-out number forms; Qwen2.5-7B compares English- and Hebrew-derived effects on Hebrew. **d**, Additional same-sentence correspondence equals observed SEEG–model similarity minus the trial-reassignment mean.

Human language processing offers a complementary perspective on shared language organization. We compare model representations with intracranial responses of 21 participants hearing the Chinese originals of the aligned materials. The shared and model-specific parts of each representation enter the same analysis of overall resemblance and additional same-sentence correspondence, the increment over whole-trial reassignment, which preserves sentence positions. We ask whether what models share also corresponds more closely to responses to the same sentences. Figure 1 brings the three comparisons together.

Our findings clarify the functional and neural significance of shared organization. **(1) Models share language relations most clearly through aligned meanings.** Relations among meanings recur most strongly across five model families and persist when meanings and language composition change. Same-language values also reconstruct local representations better than other-language values. Common relational organization therefore coexists with source-language dependence in individual representations. **(2) Recurrent language relations coexist with distributed grammatical influence and context-dependent cross-language effects.** Number edits generalize to held-out forms of marking; confined to selected units, they retain reduced effects within or below the range obtained with randomly chosen units. **(3) Overall neural resemblance and same-sentence correspondence favor different models, while shared and model-specific parts show comparable correspondence.** The two parts differ in that correspondence by less than one tenth of the lower neural reliability reference. What models share and what corresponds to the brain are therefore distinct criteria: the first concerns models alone and the second their relation to human data.

2 Language organization, grammatical influence and brain correspondence

2.1 Comparing language organization

We first ask what language organization is shared across models, comparing it in three views: relations among meanings, the arrangement of activity, and changes in response strength across successive positions in a sequence. We compare 60 languages and nine open-weight models from five families:

Llama 3, Mistral, Qwen 2.5, Gemma 2 and GPT-2. Appendix A.2 gives every model's full name, architecture and short name. The materials are aligned four-unit lexical sequences such as "friend pal treat guest".

Concept asks whether relations among aligned meanings persist across languages, using representational similarity analysis (Kriegeskorte et al., 2008); *Subspace* compares activation organization by centered-kernel alignment (CKA; Kornblith et al., 2019); *Rhythm* compares the rise and fall of response strength across stimulus positions. Units are each block's intermediate feed-forward neurons. For languages a and b, let u_a be the item dissimilarities ($1 - r$ between standardized activation-summary vectors, r the Pearson correlation), X_a the standardized item-by-feature matrix and t_a the mean absolute activation of selected units at each position, with its variability summaries. Each model M yields a pattern $d^M = (d^M_{ab})$ over every two of the 60 languages (1,770 comparisons); *recurrence* compares this pattern between models, and *reliability* one model's patterns across meanings:

$$d^{\mathrm{C}}_{ab} = 1 - r(u_a, u_b), \qquad d^{\mathrm{S}}_{ab} = 1 - \mathrm{CKA}(X_a, X_b), \qquad d^{\mathrm{R}}_{ab} = 1 - r(t_a, t_b), \tag{1}$$

$$\text{Recurrence} = \tfrac{1}{10}\textstyle\sum_{\{F,G\}} \bar{\rho}_{FG}, \qquad \text{Reliability} = \rho(d^M_f, d^M_{f'}), \quad f \neq f'. \tag{2}$$

Here ρ is the Spearman correlation, $\bar{\rho}_{FG}$ its mean over comparisons between models of families F and G, for every two of the five families $\{F, G\}$, ten in all, and d^M_f the pattern on fold f's evaluation meanings. Randomization relabels whole languages, so comparisons sharing a language stay dependent. Five folds defined over meanings separate unit selection from evaluation (Appendix A.1).

WALS structural features and Glottolog classification provide external references for these relations. The materials were translated into each language with LLM assistance, and every language version was checked by two human experts (Appendix A.3). Input changes, unit deactivation and activation replacement test dependence on material and internal components; replacement recovery measures increased similarity to the target's original representation at the same layer (Appendix B.3).

2.2 Testing the influence of grammatical number

Subject–verb agreement links subject number to a later choice, such as *is* or *are*; intervening nouns can carry conflicting number information (Bock & Miller, 1991; Franck et al., 2002; Wagers et al., 2009; Linzen et al., 2016; Gulordava et al., 2018). The *agreement margin* m is the log-probability difference between agreeing and disagreeing words; positive margins favor the agreeing form.

English materials contain 36 subject-lexeme clusters, each in 24 contexts, over regular nouns (*pilot/pilots*), irregular nouns (*child/children*), and demonstratives with invariant nouns (*this/these sheep*) (Appendices C.1–C.2). Llama-8B provided an initial evaluation; of the eight evaluated later, GPT-2-ZH did not meet the baseline agreement criterion, leaving seven (Appendix C). Each evaluation estimates the plural–singular direction v, strength α, scaling, and comparison directions from 24 clusters in two cue types, then tests the remaining type's 12 clusters, a form of marking absent from estimation. Earlier all-cluster fits, which support all three endpoints in every model, are kept separate from this held-out test (Appendices C.3–C.4). The confined edit (Equation 3) is compared with twenty random unit sets of equal size from the same layer, matched in activation strength (Appendix C.5).

The edit shifts the subject state h at the edit block to h' against its number s ($+1$ plural, -1 singular); confinement replaces v by $v_{\mathcal{U}}$, the unit-length projection $P_{\mathcal{U}}v$ onto the span of the output-projection columns of selected units $\mathcal{U}$. With m_0, m_s, m_o the margins after no edit, the subject edit and the same edit at another word, $\bar{m}_\perp$ the mean over eight equally large perpendicular subject edits, and $\langle\cdot\rangle$ the mean over a cluster's contexts, *attenuation* (A) is the loss of agreement and *selectivity* (S) its excess over both comparison edits. An opposing edit at a later word (*bridge*, b) tests compensation in the same layer; since it can improve agreement alone, *compensation* (C), evaluated with singular subjects and a conflicting intervening noun, requires joint recovery beyond the edits' separate effects, $R_v = \langle (m_{s+b} - m_s) - (m_b - m_0) \rangle$, and beyond its perpendicular-direction counterpart $R_{\text{comparison}}$:

$$h' = h - s\alpha v, \qquad v_{\mathcal{U}} = P_{\mathcal{U}}v / \|P_{\mathcal{U}}v\|_2, \tag{3}$$

$$A = \langle m_0 - m_s \rangle, \;\; S = \min\{\langle \bar{m}_\perp - m_s \rangle, \langle m_o - m_s \rangle\}, \;\; C = \min\{R_v, R_v - R_{\text{comparison}}\}. \tag{4}$$

Support for each outcome must hold in all three excluded cue types, with correction for the three outcomes within each model (Appendix C.4).

In Qwen2.5-7B (Qwen-7B), we compare the unchanged English edit with one estimated from 105 Hebrew sentence combinations on 415 other combinations from CLAMS (Mueller et al., 2020). Vo-

cabulary overlaps between estimation and evaluation; both edits act at layer 14 (zero-based). Hebrew attenuation is referenced to the perpendicular-direction edit and selectivity to the edit at another word, so the attenuation reference differs from the unedited English reference (Appendices C.8–C.9).

Both language-derived edits are tested at both original strengths, and the confined edit at the common lower strength. Sixty additional combinations cross subject and intervening-noun number while fixing distance, length, targets and edit positions, and fourteen predefined comparisons weight ten subject-word groups equally (Appendices C.10–C.13).

2.3 Comparing model and intracranial responses

Chinese sentence-listening stereoelectroencephalography (SEEG) records electrical activity through electrodes implanted for clinical care. Each of the 21 participants contributes 40 trials, each with eight analyzed four-syllable sentences of two two-character words, presented at 250 ms per syllable in the isochronous design of Ding et al. (2016). Neural responses combine low-frequency amplitude and changes in high-gamma power around word onsets. A transformation fitted on half the trials is evaluated on the other half, then exchanged; participants and model families receive equal weight (Appendices D.1–D.3). Each model's Subspace representation Z is split into the part $\hat{Z}$ predicted by ridge regression from the other model families, fitted on the training half of the trials (Appendix D.6), and the model-specific remainder $Z - \hat{Z}$; both parts enter the same held-out analysis.

Overall similarity c_{obs} is the correlation r between the off-diagonal entries of the centered model and neural similarity kernels; whole-trial reassignment π changes which model and neural sequences are compared, preserving sentence positions. Agreement between each participant and the others gives the lower neural reliability reference $\bar{r}_{\mathrm{low}}$; including that participant gives the upper reference. *Additional same-sentence correspondence* subtracts the mean $\bar{c}_\pi$ over 1,000 reassignments, and the two parts' difference η is read against the predefined margin:

$$\Delta c = c_{\mathrm{obs}} - \bar{c}_\pi, \qquad \eta = \Delta c_{\hat{Z}} - \Delta c_{Z-\hat{Z}}, \qquad |\eta| \le \bar{r}_{\mathrm{low}}/10. \tag{5}$$

Participant-bootstrap intervals quantify variation across people for fixed stimuli (Appendix D.4).

3 Related work

Shared organization across models. Models are argued to converge on shared representations (Huh et al., 2024), and one model's text representations map onto another's without paired examples (Jha et al., 2026), though similarity in vision depends on training objective and data (Ciernik et al., 2025). In multilingual models, units come to carry the same concepts across languages (Riemenschneider & Frank, 2025), and projections between language subspaces change behavior (Chang et al., 2022). Which kind of relation recurs across model families is open; accounts of language structure and ancestry (Dryer & Haspelmath, 2013; Skirgård et al., 2023a; Hammarström et al., 2026; Dunn et al., 2011; Verkerk et al., 2026) describe languages, not models, and are our external references.

Grammatical influence and intervention. Number, gender and tense often occupy shared cross-language directions; removing these features lowers concept classification accuracy to near chance (Brinkmann et al., 2025). Units sensitive to grammaticality in the residual stream causally support syntactic judgments, with overlap across agreement types (Kryvosheieva et al., 2026). Sharing varies by model and construction (Mueller et al., 2022), and transfer is greater between typologically similar languages (Boguraev et al., 2026). Whether recurrent relations coincide with concentrated grammatical influence is another question. As interpretation depends on the measure (Zhang & Nanda, 2024), we compare edits through selected and random units on held-out number forms.

Correspondence with brain responses. Encoding studies relate model representations to human responses (Schrimpf et al., 2021; Caucheteux & King, 2022; Goldstein et al., 2022). Mappings transfer across English, Chinese and French listeners hearing versions of the same narrative (Zada et al., 2025). Removing parameters important across languages reduces fMRI prediction in multilingual models (Cui et al., 2026). Data splitting and word rate contribute to apparent correspondence (Hadidi et al., 2026). Using representational similarity analysis (Kriegeskorte et al., 2008), we compare distinctions among the sentences heard. Whole-trial reassignment preserves sentence positions and gives the reference for shared and model-specific correspondence.

4 RESULTS

4.1 CONCEPT RELATIONS RECUR MOST STRONGLY ACROSS MODEL FAMILIES

Relations among meanings are what models share most consistently (Figure 2). Concept recurrence is $\rho = .422$, compared with .287 for Subspace and .039 for Rhythm; within-model consistency across meanings is higher, .660 on average (Appendix A.5). This relational commonality persists across models whose distributed activity and sequential response profiles differ.

This shared organization extends beyond identical content and the largest sampled language family. Concept recurrence remains .313 when models are evaluated on different meanings and .376 after excluding Indo-European languages. Selected units exceed equally sized random sets in every fold (Appendix A.5). The recurrence therefore extends across evaluation meanings and language composition. The next comparison is whether the units selected for these relations also concentrate grammatical influence (Section 4.3). Aggregating languages into families changes which representation tracks documented structure, with Rhythm closest to Glottolog and Subspace closest to WALS at that scale (Appendix A.4).

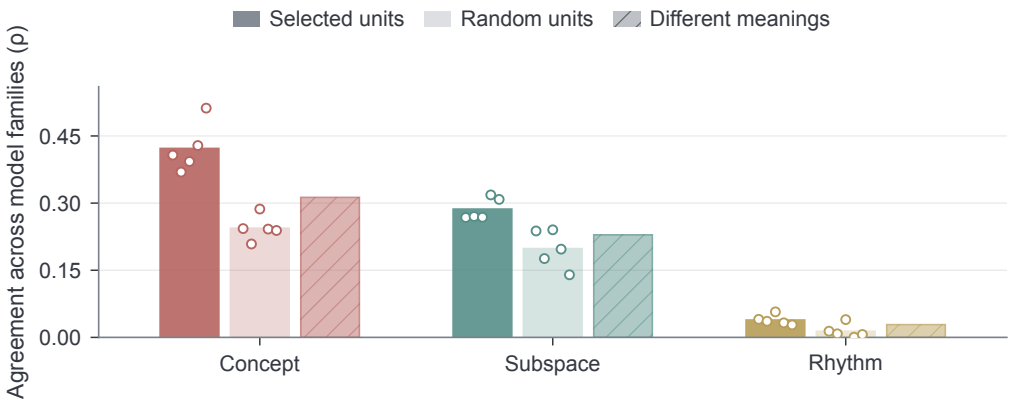

Figure 2: **Relations among meanings recur most strongly across model families.** Bars compare language-distance patterns, weighting each family-to-family comparison equally. Selected and equally sized random unit sets use the same held-out meanings; the third comparison uses different meanings. Points show the five overlapping folds; the different-meaning estimate pools them.

The external-reference analysis relates model language distances to documented structure. After adjustment for genealogy, geography, writing system and material differences, the largest mean WALS coefficient is .0457 for Concept (Figure S1; Appendix A.3). The Llama 3 and Gemma 2 families assign more similar Rhythm representations to languages sharing present-day grammatical feature combinations supported by evolutionary analyses (Appendix A.6). The recurrent relation is thus a reproducible property of the models themselves, beyond what the structural and historical descriptors tested here capture.

4.2 SOURCE LANGUAGE CONDITIONS LOCAL RECONSTRUCTION

The three representations respond differently to changes in the same linguistic material. Scrambling the input most strongly changes concept relations, whereas swapping the middle two aligned units affects Rhythm more than replacing a phrase does; Subspace remains comparatively stable (Figure S2; Appendices B and B.1). Preserved overall activation organization can therefore mask changes in relations among meanings and in sequential response strength.

Replacing part of a representation with values from the same language better restores its original organization than replacement from another language (Figure 3). This contrast appears in every model's mean and is largest for Concept, with similarity changes of $+.105$ and $-.139$, respectively. Subspace and Rhythm show the same average direction, with variation across individual unit selections (Figure S3). Recovery is measured against each target's own original organization at that layer.

Common relations across languages therefore coexist with source-dependent reconstruction of a particular representation.

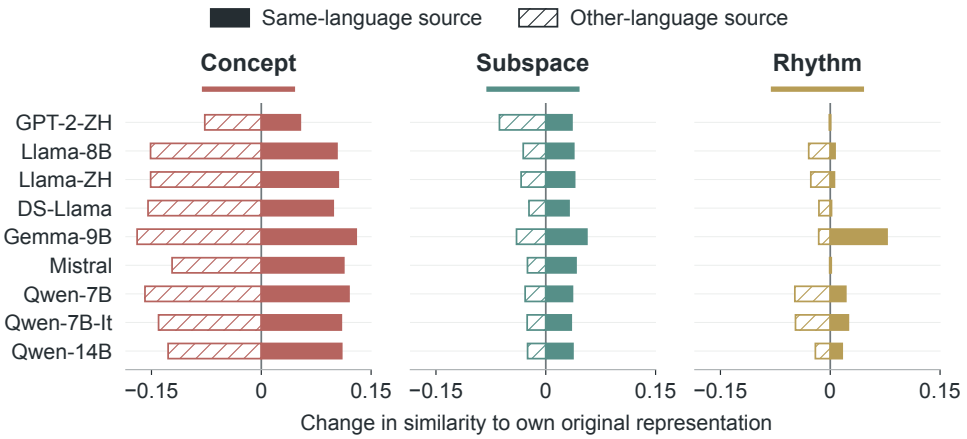

Figure 3: **Source language conditions local reconstruction.** Bars show all nine model means across Concept, Subspace and Rhythm on a common scale. Positive changes indicate recovery toward each recipient's original same-layer representation. Solid and hatched bars distinguish same- and other-language sources. Individual unit selections vary (Figure S3).

The behavioral effects of broad representation edits vary across tasks and the type of change applied (Appendix B.4). Subject–verb agreement gives the next comparison a precise functional question: does changing subject-number information weaken preference for the agreeing verb more than comparable changes in another direction or at another word, and do the units that carry the shared relations concentrate that influence?

4.3 Number information affects agreement across different forms of marking

Represented number information influences agreement beyond the forms used to estimate it: a number change estimated from two forms of number marking also affects agreement when number is expressed in the third. Across regular nouns, irregular nouns, and demonstratives with invariant heads, agreement is weakened more by the number edit than by either comparison edit in all seven models (Figure 4; Appendix C.4). Compensation through a later-position change meets the same cross-cue requirement in six models, so the intervention both weakens agreement for the excluded form and supports recovery through a later edit.

Recurrent language relations coexist with distributed grammatical influence. Confined to the selected units' output space at the edit block, the number edit keeps its fitted strength and length but retains 3–41% of its attenuation and selectivity, and it still weakens agreement across all three forms of marking in five models. Against twenty sets of random units, the confined edit lies within the random range in five models and below at least eighteen of the twenty sets in Llama-ZH and DS-Llama; its compensation lies within the range in four models and below at least eighteen sets in the other three (Figure 4; Appendix C.5). Recurrent relations and grammatical influence are therefore distinct properties of one representation.

Mistral shows why testing the later edit alone matters: for irregular nouns, editing the bridge alone improves agreement almost as much as editing both positions, so the recovery there arises chiefly from the later edit itself (Figure S4; Appendix C.4).

Gemma-9B shows that selective grammatical influence can depend on surrounding noun-number information in opposite ways. A mismatching intervening noun makes the edit more disruptive for regular and irregular forms, but less disruptive in every demonstrative cluster (Figure S4; Appendices C.6–C.7). All three cue groups retain attenuation, selectivity, and compensation; in the earlier all-cluster test, context amplifies disruption in six models, whereas Gemma-9B's ratio lies inside the prespecified equivalence region (Appendix C.7). An average close to zero can therefore hide opposite context effects.

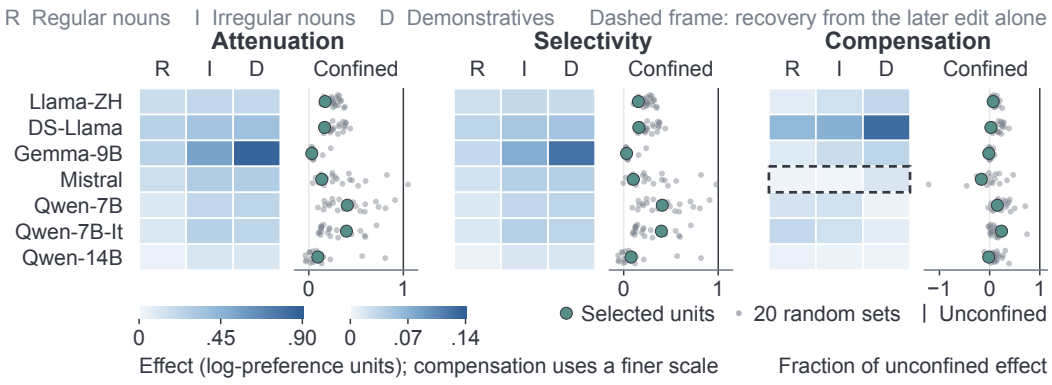

Figure 4: **Number edits selectively weaken agreement for every excluded form of marking.** Color gives mean effects for each excluded number form (12 lexical clusters); attenuation and selectivity share a scale; compensation has a finer scale. Mistral's dashed frame marks compensation whose recovery arises chiefly from the later edit alone (Appendix C.4). Dots give the edit confined to the selected units (teal) or to twenty sets of random units (gray), as a fraction of the unconfined effect, marked by the line at 1.

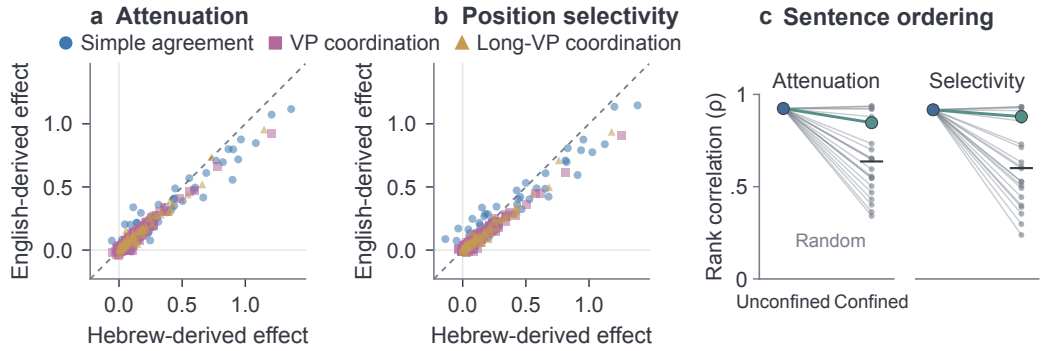

Figure 5: **English- and Hebrew-derived edits produce similar sentence rankings that survive confinement of the edit. a,b**, Points compare raw effects in log-preference units across 415 Hebrew configurations at the common lower strength, averaging reciprocal number items; colors and symbols distinguish constructions, and the diagonal marks equal effects. Vocabulary overlaps with estimation (Appendix C.8). **c**, Rank correlation with the Hebrew-derived edit: unconfined edit, then confined to 398 selected units (teal) or to twenty sets of random units (gray; tick, their mean).

4.4 English and Hebrew edits affect the same sentences most strongly

Sentences more affected by the Hebrew-derived number change are more affected by the English-derived change. At the common lower strength, Spearman correlations between raw sentence effects exceed .91 for both attenuation and selectivity (Figure 5). This relation holds across constructions whose average effects differ. The Hebrew-derived edit moves agreement in the intended direction in 409 of 415 configurations (Appendices C.8–C.9). The shared ordering thus identifies similar sentence susceptibility across effects of different magnitude.

The sentence ordering survives confinement of the edit to the 398 selected units of Qwen-7B: at the common lower strength its Hebrew effects correlate at $\rho = .847$ with the Hebrew-derived edit (.924 unconfined; .34–.94, mean .64, for twenty random unit sets), while its mean effect falls to .50 of the Hebrew-derived effect (.87 unconfined; Figure 5c). Sentence rankings remain similar across language-derived edits and after confinement, while confinement reduces effect size.

The magnitude of transfer depends on the size of the intervention. At the common lower strength, the English-derived edit retains about 87% of attenuation and 82% of selectivity; increasing both edits

enlarges the Hebrew-derived advantage (Figure 6; Appendix C.10). The original unequal-strength comparison uses normalized responses: the English edit retains about 30% of the Hebrew effect, with a rank correlation near .88. Ordering and magnitude thus generalize differently.

Effects of edits at another word also correspond between the two operations ($\rho = .93$), whereas alternative-direction effects correspond at $\rho = -.05$, so comparing both controls separates general susceptibility from the influence specific to the number direction (Appendix C.11).

4.5 Surrounding number information changes susceptibility to the edit

The English-derived edit has a stronger effect on Hebrew agreement when the intervening noun carries conflicting number. Changing that noun's inflected form increases attenuation and selectivity at the lower strength by about .06 log-preference units; the Hebrew-derived increase is about .07 (Figure 6). Noun number and token identity change together; edit direction, strength, distance, targets and edit positions are fixed. Surrounding number information therefore conditions susceptibility; all four contrasts survive the predefined 14-comparison correction.

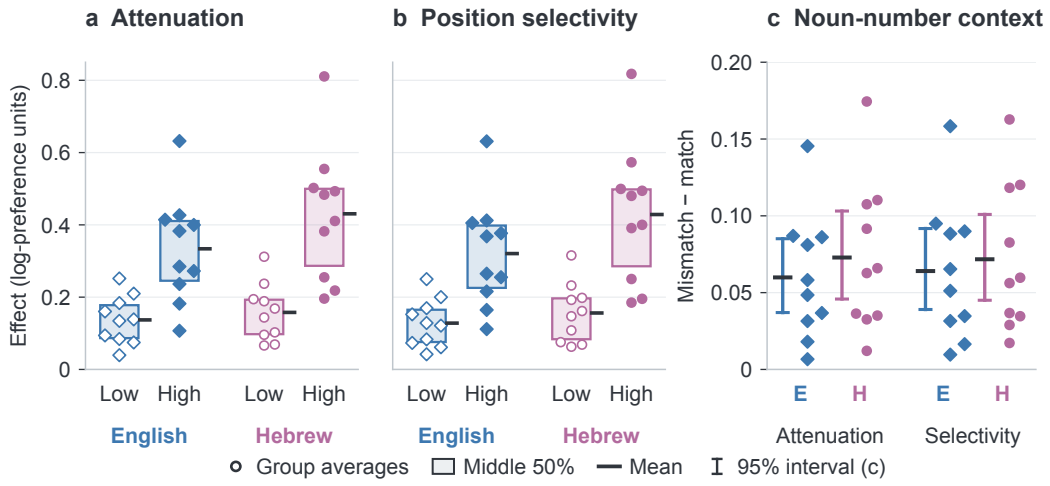

Figure 6: **Strength and noun-number context shape grammatical influence. a,b**, Points show the same ten subject-word group means at lower/higher strengths (8.55/13.98); boxes span the middle 50%, and dark ticks mark overall means. **c**, Group mismatch-minus-match effects at the lower strength, with means and marginal 95% intervals. Colors and symbols distinguish English- and Hebrew-derived edits in Qwen-7B.

Intervention susceptibility also separates from the original difficulty of agreement. Conflicting nouns reduce the unedited agreement preference more for plural subjects, but amplify the edit more for singular subjects, so initial preference and sensitivity to changed number information capture different properties (Appendix C.12).

4.6 Overall similarity and same-sentence correspondence rank models differently

We ask whether shared and model-specific representations differ in brain correspondence, beginning with model rankings. Overall brain–model similarity and additional same-sentence correspondence, computed on model Subspace and integrated Chinese intracranial responses, rank the models differently (Figure 7). Gemma-9B has the highest overall similarity (.0393) and an increment of +.00009; DS-Llama and GPT-2-ZH combine lower overall scores with larger increments (Figure 7b).

Most overall correspondence survives whole-trial reassignment: the increment is .00189 within the observed family-weighted correlation of .0185. Its positive participant-bootstrap interval describes the average across people for the fixed stimuli; the trial-reassignment test (Figure 7c; Appendix D.4) gives $p = .072$ for stimulus specificity.

The increment is .72% of the lower interparticipant reliability reference (.00189 divided by .264), with a 90% interval of .14–1.32%. Omitting any one participant, model family or trial preserves the positive average (Appendix D.5). Overall neural resemblance and correspondence to the same sentences therefore capture different properties of a representation.

Shared and model-specific parts have comparable same-sentence correspondence within the predefined margin. For the part predicted from other model families and the model-specific remainder, the means are .00236 and .00060, with families weighted equally. Their difference is .00176 (95% participant interval $[-.00163, .00526]$; reassignment $p = .137$). Its 90% interval $[-.00112, .00471]$ falls within $\pm.0264$, one tenth of the lower neural reliability reference. Point estimates stay within this margin when any participant or model family is omitted and for low-frequency and high-gamma responses considered separately (Figure 7d; Appendix D.6). Sharing across models and correspondence with human language processing therefore provide distinct criteria for understanding representations in multilingual models.

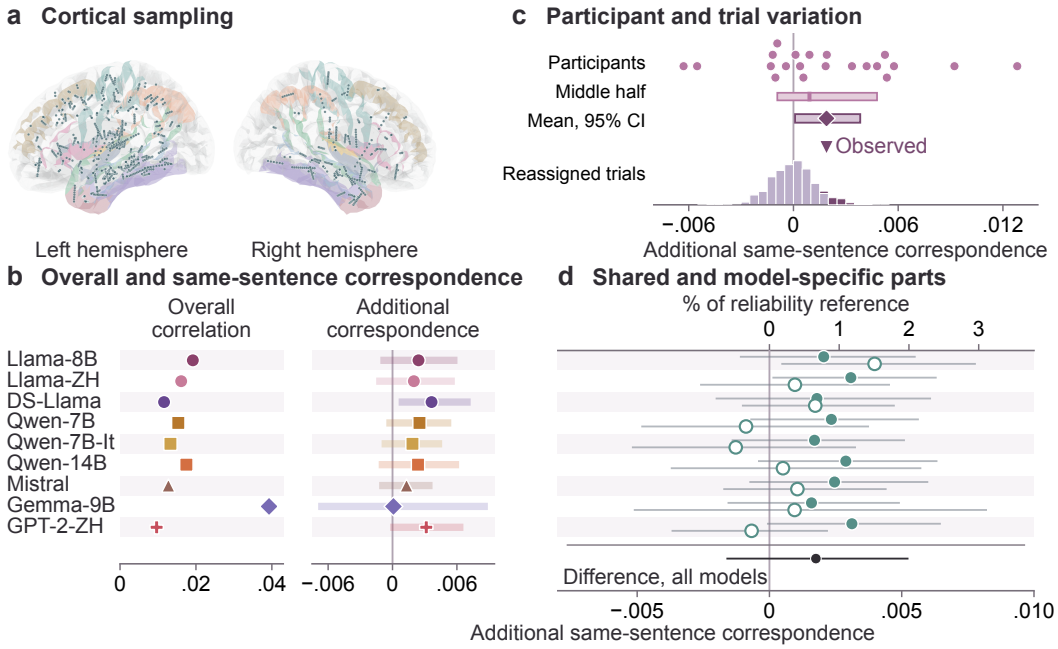

Figure 7: **Overall and same-sentence correspondence rank models differently in Chinese SEEG.** **a**, Coordinates cover 20 of 21 participants. **b**, Overall correlations and increments with unadjusted participant 95% intervals. **c**, Participant increments, middle-half range, median and mean 95% interval; reassigned-trial histogram centered on its mean, with observed increment ($p = .072$). **d**, Shared (filled) and model-specific (open) parts of each model's representation, rows as in **b**, with participant 95% intervals, and their family-weighted difference; on the upper axis, one tenth of the reliability reference is 10.

5 Conclusion

Multilingual models combine common language organization with variation in how linguistic information affects behavior. Recurring meaning relations accompany source-dependent local reconstruction and grammatical influence that extends across number forms and beyond the selected units. Cross-language transfer preserves similar sentence sensitivities, with effect sizes shaped by edit strength and number context. Intracranial responses add a complementary criterion: overall resemblance and same-sentence correspondence favor different models, while shared and model-specific parts have comparable correspondence within the predefined reliability margin. By clarifying how multilingual representations relate to cross-language computation and human language processing, this work informs theories of multilingual intelligence that account for cross-language generalization while accommodating the diverse computational forms of linguistic knowledge.

AI USE STATEMENT

In this work, we used Claude Opus 4.6 to assist with translating the aligned lexical materials into the languages of the study; every language version was checked by two human experts. Additionally, ChatGPT was used for code assistance and language editing. We have reviewed all AI-assisted work and take responsibility for the final content of this work.

ETHICS STATEMENT

The SEEG acquisition protocol received institutional review board approval, and informed consent was obtained before data collection. Participants received oral task instructions and information about auditory-stimulation risks; compensation was prorated to task completion. Recordings were obtained from patients undergoing clinical evaluation for epilepsy. Clinical signals, precise electrode coordinates, and direct identifiers are excluded from the anonymous release. Institution-identifying approval and source details are withheld for double-blind review; clinical data access remains subject to the acquisition protocol's governance.

REPRODUCIBILITY STATEMENT

Code, results, source data, and experimental settings are available at `https://anonymous.4open.science/r/Multilingual-69FC/`; the release documents how each analysis is reproduced from these inputs, including the model inference it requires, and preserves signed results and their normalization choices. Appendices A–D describe sampling units, outcomes and comparisons; Appendices D.1–D.3 detail SEEG acquisition, the 21-participant sample, signal processing, trial-based evaluation and anatomical coverage available for 20 participants. Non-identifying neural summaries support reproduction of aggregate statistics. Repeating neural preprocessing, trial randomization and reliability estimation requires restricted clinical recordings; rebuilding the anatomical coverage additionally requires precise electrode coordinates. These clinical inputs are excluded from the anonymous release.

REFERENCES

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. On the cross-lingual transferability of monolingual representations. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4623–4637, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.421. URL https://aclanthology.org/2020.acl-main.421/.

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. The Belebele benchmark: a parallel reading comprehension dataset in 122 language variants. 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. 749–775, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.44. URL https://aclanthology.org/2024.acl-long.44/.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, March 2022. doi: 10.1162/coli_a_00422. URL https://aclanthology.org/2022.cl-1.7/.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 57(1):289–300, January 1995. ISSN 1467-9868. doi: 10.1111/j.2517-6161.1995.tb02031.x. URL https://doi.org/10.1111/j.2517-6161.1995.tb02031.x.

Kathryn Bock and Carol A Miller. Broken agreement. *Cognitive Psychology*, 23(1):45–93, 1991. doi: 10.1016/0010-0285(91)90003-7. URL https://doi.org/10.1016/0010-0285(91)90003-7.

Sasha Boguraev, Toshiki Nakai, Kyle Mahowald, and Julius Steuer. Causal interventions reveal typologically organized syntactic mechanisms in multilingual language models. arXiv preprint arXiv:2608.28924, 2026. URL https://arxiv.org/abs/2608.28924.

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/.

Charlotte Caucheteux and Jean-Rémi King. Brains and algorithms partially converge in natural language processing. *Communications Biology*, 5(1):134, 2022. doi: 10.1038/s42003-022-03036-1. URL https://doi.org/10.1038/s42003-022-03036-1.

Tyler A. Chang, Zhuowen Tu, and Benjamin K. Bergen. The geometry of multilingual language model representations. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 119–136, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.9. URL https://aclanthology.org/2022.emnlp-main.9/.

Laure Ciernik, Lorenz Linhardt, Marco Morik, Jonas Dippel, Simon Kornblith, and Lukas Muttenthaler. Objective drives the consistency of representational similarity across datasets. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 10920–10948. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/ciernik25a.html.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. XNLI: Evaluating cross-lingual sentence representations. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2475–2485, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1269. URL https://aclanthology.org/D18-1269/.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 8440–8451, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.747. URL https://aclanthology.org/2020.acl-main.747/.

Yang Cui, Jingyuan Sun, Yizheng Sun, Yifan Wang, Yunhao Zhang, Jixing Li, Shaonan Wang, Hongpeng Zhou, John Hale, Chengqing Zong, and Goran Nenadic. Computational lesions in multilingual language models separate shared and language-specific brain alignment. arXiv preprint arXiv:2604.10627, 2026. URL https://arxiv.org/abs/2604.10627.

Yiming Cui and Xin Yao. Rethinking LLM language adaptation: A case study on Chinese Mixtral. *arXiv preprint arXiv:2403.01851*, 2024. URL https://arxiv.org/abs/2403.01851.

Yiming Cui, Ziqing Yang, and Xin Yao. Efficient and effective text encoding for Chinese LLaMA and Alpaca. *arXiv preprint arXiv:2304.08177*, 2023. URL https://arxiv.org/abs/2304.08177.

Nai Ding, Lucia Melloni, Hang Zhang, Xing Tian, and David Poeppel. Cortical tracking of hierarchical linguistic structures in connected speech. *Nature Neuroscience*, 19(1):158–164, 2016. doi: 10.1038/nn.4186. URL https://doi.org/10.1038/nn.4186.

Matthew S. Dryer and Martin Haspelmath (eds.). *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig, 2013. URL https://wals.info/.

Michael Dunn, Simon J. Greenhill, Stephen C. Levinson, and Russell D. Gray. Evolved structure of language shows lineage-specific trends in word-order universals. *Nature*, 473(7345): 79–82, 2011. doi: 10.1038/nature09923. URL https://www.nature.com/articles/nature09923.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175, 2021. doi: 10.1162/tacl_a_00359. URL https://aclanthology.org/2021.tacl-1.10/.

Julie Franck, Gabriella Vigliocco, and Janet Nicol. Subject-verb agreement errors in French and English: The role of syntactic hierarchy. *Language and Cognitive Processes*, 17(4): 371–404, 2002. doi: 10.1080/01690960143000254. URL https://doi.org/10.1080/01690960143000254.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 9574–9586. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/4f5c422f4d49a5a807eda27434231040-Paper.pdf.

Atticus Geiger, Zhengxuan Wu, Christopher Potts, Thomas Icard, and Noah Goodman. Finding alignments between interpretable causal variables and distributed neural representations. In Francesco Locatello and Vanessa Didelez (eds.), *Proceedings of the Third Conference on Causal Learning and Reasoning*, volume 236 of *Proceedings of Machine Learning Research*, pp. 160–187. PMLR, 01–03 Apr 2024. URL https://proceedings.mlr.press/v236/geiger24a.html.

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

Ariel Goldstein, Zaid Zada, Eliav Buchnik, Mariano Schain, Amy Price, Bobbi Aubrey, Samuel A. Nastase, Amir Feder, Dotan Emanuel, Alon Cohen, Aren Jansen, Harshvardhan Gazula, Gina Choe, Aditi Rao, Catherine Kim, Colton Casto, Lora Fanda, Werner Doyle, Daniel Friedman, Patricia Dugan, Lucia Melloni, Roi Reichart, Sasha Devore, Adeen Flinker, Liat Hasenfratz, Omer Levy, Avinatan Hassidim, Michael Brenner, Yossi Matias, Kenneth A. Norman, Orrin Devinsky, and Uri Hasson. Shared computational principles for language processing in humans and deep language models. *Nature Neuroscience*, 25(3):369–380, 2022. doi: 10.1038/s41593-022-01026-4. URL https://doi.org/10.1038/s41593-022-01026-4.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, et al. The Llama 3 herd of models. arXiv preprint arXiv:2407.21783, 2024. URL https://arxiv.org/abs/2407.21783.

Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. Colorless green recurrent networks dream hierarchically. In Marilyn Walker, Heng Ji, and Amanda Stent (eds.), *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 1195–1205, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1108. URL https://aclanthology.org/N18-1108/.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645(8081):633–638, 2025. doi: 10.1038/s41586-025-09422-z. URL https://www.nature.com/articles/s41586-025-09422-z.

Nima Hadidi, Ebrahim Feghhi, Bryan H. Song, Idan A. Blank, and Jonathan C. Kao. Spurious alignment between large language models and brains can emerge from non-robust methods and overlooked confounds. *Nature Communications*, 17(1):5769, 2026. doi: 10.1038/s41467-026-72253-7. URL https://doi.org/10.1038/s41467-026-72253-7.

Harald Hammarström, Robert Forkel, Martin Haspelmath, and Sebastian Bank. *Glottolog 5.3*. Max Planck Institute for Evolutionary Anthropology, Leipzig, 2026. doi: 10.5281/zenodo.18840935. URL https://doi.org/10.5281/zenodo.18840935.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2733–2743, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1275. URL https://aclanthology.org/D19-1275/.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979. URL https://www.jstor.org/stable/4615733.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. Position: The Platonic representation hypothesis. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 20617–20642. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/huh24a.html.

Rishi Jha, Collin Zhang, Vitaly Shmatikov, and John Morris. Harnessing the universal geometry of embeddings. *Advances in Neural Information Processing Systems*, 38:45963–45987, 2026. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/4175dee33d6145cb8f0323703d138a53-Abstract-Conference.html.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3519–3529. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/kornblith19a.html.

Nikolaus Kriegeskorte, Marieke Mur, and Peter Bandettini. Representational similarity analysis – connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2:4, 2008. doi: 10.3389/neuro.06.004.2008. URL https://doi.org/10.3389/neuro.06.004.2008.

Daria Kryvosheieva, Andrea Gregor de Varda, Evelina Fedorenko, and Greta Tuckute. Different types of syntactic agreement recruit the same units within large language models. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 209–227, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.7. URL https://aclanthology.org/2026.acl-long.7/.

Patrick Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. MLQA: Evaluating cross-lingual extractive question answering. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7315–7330, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.653. URL https://aclanthology.org/2020.acl-main.653/.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535, 2016. doi: 10.1162/tacl_a_00115. URL https://aclanthology.org/Q16-1037/.

Chaitanya Malaviya, Graham Neubig, and Patrick Littell. Learning language representations for typology prediction. In Martha Palmer, Rebecca Hwa, and Sebastian Riedel (eds.), *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 2529–2535, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1268. URL https://aclanthology.org/D17-1268/.

Aaron Mueller, Garrett Nicolai, Panayiota Petrou-Zeniou, Natalia Talmina, and Tal Linzen. Cross-linguistic syntactic evaluation of word prediction models. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5523–5539, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.490. URL https://aclanthology.org/2020.acl-main.490/.

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/.

Telmo Pires, Eva Schlinger, and Dan Garrette. How multilingual is multilingual BERT? In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4996–5001, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1493. URL https://aclanthology.org/P19-1493/.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. Technical report, OpenAI, 2019. URL https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

Taraka Rama, Lisa Beinborn, and Steffen Eger. Probing multilingual BERT for genetic and typological signals. In Donia Scott, Nuria Bel, and Chengqing Zong (eds.), *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 1214–1228, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.105. URL https://aclanthology.org/2020.coling-main.105/.

Frederick Riemenschneider and Anette Frank. Cross-lingual generalization and compression: From language-specific to shared neurons. 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. 13470–13491, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.661. URL https://aclanthology.org/2025.acl-long.661/.

Martin Schrimpf, Idan Asher Blank, Greta Tuckute, Carina Kauf, Eghbal A. Hosseini, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, 118(45):e2105646118, 2021. doi: 10.1073/pnas.2105646118. URL https://doi.org/10.1073/pnas.2105646118.

Hedvig Skirgård, Hannah J. Haynie, Damián E. Blasi, Harald Hammarström, Jeremy Collins, Jay J. Latarche, Jakob Lesage, Tobias Weber, Alena Witzlack-Makarevich, Sam Passmore, Angela Chira, Luke Maurits, Russell Dinnage, Michael Dunn, Ger Reesink, Ruth Singer, Claire Bowern, Patience Epps, Jane Hill, Outi Vesakoski, Martine Robbeets, Noor Karolin Abbas, Daniel Auer, Nancy A. Bakker, Giulia Barbos, Robert D. Borges, Swintha Danielsen, Luise Dorenbusch, Ella Dorn, John Elliott, Giada Falcone, Jana Fischer, Yustinus Ghanggo Ate, Hannah Gibson, Hans-Philipp Göbel, Jemima A. Goodall, Victoria Gruner, Andrew Harvey, Rebekah Hayes, Leonard Heer, Roberto E. Herrera Miranda, Nataliia Hübler, Biu Huntington-Rainey, Jessica K. Ivani, Marilen Johns, Erika Just, Eri Kashima, Carolina Kipf, Janina V. Klingenberg, Nikita König, Aikaterina Koti, Richard G. A. Kowalik, Olga Krasnoukhova, Nora L. M. Lindvall, Mandy Lorenzen, Hannah Lutzenberger, Tânia R. A. Martins, Celia Mata German, Suzanne van der Meer, Jaime Montoya Samamé, Michael Müller, Saliha Muradoglu, Kelsey Neely, Johanna Nickel, Miina Norvik, Cheryl Akinyi Oluoch, Jesse Peacock, India O. C. Pearey, Naomi Peck, Stephanie Petit, Sören Pieper, Mariana Poblete, Daniel Prestipino, Linda Raabe, Amna Raja, Janis Reimringer, Sydney C. Rey, Julia Rizaew, Eloisa Ruppert, Kim K. Salmon, Jill Sammet, Rhiannon Schembri, Lars Schlabbach, Frederick W. P. Schmidt, Amalia Skilton, Wikaliler Daniel Smith, Hilário de Sousa, Kristin Sverredal, Daniel Valle, Javier Vera, Judith Voß, Tim Witte, Henry Wu, Stephanie Yam, Jingting Ye, Maisie Yong, Tessa Yuditha, Roberto Zariquiey, Robert Forkel, Nicholas Evans, Stephen C. Levinson, Martin

Haspelmath, Simon J. Greenhill, Quentin D. Atkinson, and Russell D. Gray. Grambank reveals the importance of genealogical constraints on linguistic diversity and highlights the impact of language loss. *Science Advances*, 9(16):eadg6175, April 2023a. doi: 10.1126/sciadv.adg6175. URL https://doi.org/10.1126/sciadv.adg6175.

Hedvig Skirgård, Hannah J. Haynie, Harald Hammarström, Russell Barlow, Damián E. Blasi, Jeremy Collins, Jay Latarche, Jakob Lesage, Tobias Weber, Alena Witzlack-Makarevich, Michael Dunn, Ger Reesink, Ruth Singer, Johannes Englisch, Angela Chira, Annemarie Verkerk, Russell Dinnage, Luke Maurits, Sam Passmore, Claire Bowern, Patience Epps, Jane Hill, Outi Vesakoski, Noor Karolin Abbas, Sunny Ananth, Daniel Auer, Nancy A. Bakker, Giulia Barbos, Anina Bolls, Robert D. Borges, Mitchell Browen, Hoju Cha, Lennart Chevallier, Swintha Danielsen, Hugo de Vos, Sinoël Dohlen, Luise Dorenbusch, Ella Dorn, Marie Duhamel, Farah El Haj Ali, John Elliott, Grace Ephraums, Giada Falcone, Anna-Maria Fehn, Jana Fischer, Yustinus Ghanggo Ate, Hannah Gibson, Hans-Philipp Göbel, Jemima A. Goodall, Samuel Griggs, Victoria Gruner, Andrew Harvey, Rebekah Hayes, Leonard Heer, Roberto E. Herrera Miranda, Nataliia Hübler, Biu H. Huntington-Rainey, Guglielmo Inglese, Jessica K. Ivani, Marilen Johns, Erika Just, Ivan Kapitonov, Eri Kashima, Carolina Kipf, Janina V. Klingenberg, Nikita König, Aikaterina Koti, Richard G. A. Kowalik, Olga Krasnoukhova, Kate Lynn Lindsey, Nora L. M. Lindvall, Mandy Lorenzen, Hannah Lutzenberger, Manuel Rüdisühli, Alexandra Marley, Tânia R. A. Martins, Marvin Leonard Martiny, Celia Mata German, Suzanne van der Meer, Jacob Menschel, Jaime Montoya, Michael Müller, Saliha Muradoglu, HunterGatherer, David Nash, Kelsey Neely, Johanna Nickel, Miina Norvik, Olga Olina, Bruno Olsson, Cheryl Akinyi Oluoch, David Osgarby, Leah Pappas, Jesse Peacock, India O.C. Pearey, Naomi Peck, Jana Peter, Stephanie Petit, Sören Pieper, Mariana Poblete, Daniel Prestipino, Linda Raabe, Amna Raja, Tihomir Rangelov, Janis Reimringer, Sydney C. Rey, Julia Rizaew, Kristian Roncero Toledo, Eloisa Ruppert, Kim K. Salmon, Jill Sammet, Rhiannon Schembri, Lars Schlabbach, Frederick W. P. Schmidt, Dineke Schokkin, Jeff Siegel, Jane Simpson, Amalia Skilton, Hilário de Sousa, Kristin Sverredal, Daniel Valle, Javier Vera, Judith Voß, Daniel Wikalier Smith, Nicholas Williams, Tim Witte, Henry Wu, Stephanie Yam, Jingting Ye, Maisie Yong, Tessa Yuditha, Roberto Zariquiey, Olena Shcherbakova, Robert Forkel, Nicholas Evans, Stephen C. Levinson, Martin Haspelmath, Simon J. Greenhill, Quentin D. Atkinson, and Russell D. Gray. Grambank v1.0 (v1.0.3) [Data set]. Zenodo, 2023b. URL https://doi.org/10.5281/zenodo.7844558.

Annemarie Verkerk, Olena Shcherbakova, Hannah J. Haynie, Hedvig Skirgård, Christoph Rzymski, Quentin D. Atkinson, Simon J. Greenhill, and Russell D. Gray. Enduring constraints on grammar revealed by Bayesian spatiophylogenetic analyses. *Nature Human Behaviour*, 10(1):126–136, 2026. doi: 10.1038/s41562-025-02325-z. URL https://www.nature.com/articles/s41562-025-02325-z.

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.

Matthew W. Wagers, Ellen F. Lau, and Colin Phillips. Agreement attraction in comprehension: Representations and processes. *Journal of Memory and Language*, 61(2):206–237, 2009. doi: 10.1016/j.jml.2009.04.002. URL https://doi.org/10.1016/j.jml.2009.04.002.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, et al. Qwen2.5 technical report. arXiv preprint arXiv:2412.15115, 2024. URL https://arxiv.org/abs/2412.15115.

Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. PAWS-X: A cross-lingual adversarial dataset for paraphrase identification. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3687–3692, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1382. URL https://aclanthology.org/D19-1382/.

Zaid Zada, Samuel A Nastase, Jixing Li, and Uri Hasson. Brains and language models converge on a shared conceptual space across different languages. arXiv preprint arXiv:2506.20489, 2025. URL https://arxiv.org/abs/2506.20489.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *International Conference on Learning Representations*, volume 2024, pp. 1651–1678, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/06a52a54c8ee03cd86771136bc91eb1f-Abstract-Conference.html.

Zhe Zhao, Hui Chen, Jinbin Zhang, Xin Zhao, Tao Liu, Wei Lu, Xi Chen, Haotang Deng, Qi Ju, and Xiaoyong Du. UER: An open-source toolkit for pre-training models. In Sebastian Padó and Ruihong Huang (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, pp. 241–246, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-3041. URL https://aclanthology.org/D19-3041/.

A MULTILINGUAL RELATIONS AND LINGUISTIC REFERENCES

A.1 LANGUAGES, MATERIALS, AND REPRESENTATION MEASURES

The 60-language comparison asks which relations recur across nine models from five families (Table S1; Appendix A.2). Each language contributes aligned meanings and positions, and every two of the 60 languages are compared, 1,770 comparisons in all. Because each language enters 59 of these comparisons, language-label permutations, language/family deletion analyses, and dyadic models preserve the dependence among them.

Aligned two-unit lexical groups are combined into four-unit sequences, such as "friend pal treat guest" and "car auto turn bend". Spaces mark alignment units; underscores join multiword expressions within a unit. The 320 occurrences from 40 trials comprise 100 unique four-unit groups, with all repetitions of a group assigned to one fold. This restricted vocabulary holds meanings and positions comparable across languages. The materials were translated into each language with LLM assistance, and every language version was checked by two human experts.

The three measures locate shared organization at different levels. Units are the intermediate neurons of each block's feed-forward sublayer. For Concept, each aligned item is represented by activation summaries: mean, mean absolute activation and its standard deviation. Summaries are computed over semantic candidate units in each layer, by meaning category and overall, and each summary is standardized across items. Item dissimilarity is one minus the Pearson correlation between summary vectors. Language distance is one minus the Pearson correlation between item-dissimilarity vectors, asking whether relations among meanings persist across languages.

Subspace distance is one minus linear centred-kernel alignment between the two languages' item-by-feature matrices, with features standardized across items. Features comprise the same layerwise activation summaries over the integrated syntactic and semantic candidates and over the syntactic, semantic and intersection subsets. This measure compares overall activation organization. Rhythm distance is one minus the correlation between sequential strength profiles. These profiles combine mean absolute activation across selected units at each position, averaged across items, with profile-variability summaries.

Recurrence asks which of these language comparisons agrees most consistently across models. Unit selection uses discovery meanings only. Evaluation crosses unit classes, shallow/middle/deep layer tiers and half/matched/double counts, with equally sized random units per layer. These prespecified tiers are aggregated independently of the block used for number intervention.

A.2 MODEL SET AND ARCHITECTURE

The nine models are open-weight decoder-only transformers from five families (Table S2). Three share the Llama 3 architecture: Llama-8B is the released base model, DS-Llama is the DeepSeek-R1 reasoning-distilled model whose base model is Llama-3.1-8B, and Llama-ZH continues the pretraining of Meta-Llama-3-8B on 120 GB of Chinese text. Qwen-7B-It is the instruction-tuned release of Qwen-7B, and Qwen-14B is the larger base model of the same family. Mistral, Gemma-9B and GPT-2-ZH each represent their family alone; GPT-2-ZH is the 36-layer GPT-2 model pretrained on Chinese text with UER-py. Family membership defines the ten family-to-family comparisons that receive equal weight in the recurrence analysis (Appendix A.1). The models are the public releases Meta-Llama-3-8B, DeepSeek-R1-Distill-Llama-8B, hfl/llama-3-chinese-8b, Mistral-Nemo-Base-2407, Qwen2.5-7B, Qwen2.5-7B-Instruct, Qwen2.5-14B, gemma-2-9b and uer/gpt2-large-chinese-cluecorpussmall; Mistral-Nemo-Base-2407 was released by Mistral AI with NVIDIA in July 2024. Llama-3-Chinese-8B is a model of the Chinese-LLaMA-Alpaca project, which adapts open models to Chinese by continued pretraining on Chinese text, with vocabulary extension introduced for LLaMA (Cui et al., 2023) and its necessity re-examined for a later base model (Cui & Yao, 2024); it is a Chinese language-adaptation version based on Llama 3 (Grattafiori et al., 2024). Parameter counts are the stored weights of each model, and the architecture values are read from each model's configuration. The intervention experiments edit the block at half depth, listed as the zero-based edit block; GPT-2-ZH did not meet the baseline agreement criterion and was not intervened on (Appendix C). Table S3 ties each model's stored representation of the 320 sentence occurrences used in the brain comparison, and the units selected at its edit block for the confined edit, to these releases.

Table S1: Languages in the common representation comparison. Identifiers preserve the linguistic resources' source spellings; the online inventory additionally gives WALS codes and observed-feature coverage.

Language	ISO	Glottocode	Language	ISO	Glottocode
Mandarin	cmn	mand1415	Maori	mri	maor1246
English	eng	stan1293	Khmer	khm	cent1989
Korean	kor	kore1280	Burmese	mya	nucl1310
Japanese	jpn	nucl1643	Finnish	fin	finn1318
Spanish	spa	stan1288	Hungarian	hun	hung1274
French	fra	stan1290	Georgian	kat	nucl1302
German	deu	stan1295	Guaraní	gug	para1311
Hindi	hin	hind1269	Navajo	nav	nava1243
Arabic (Modern Standard)	arb	stan1318	Greenlandic (West)	kal	kala1399
Russian	rus	russ1263	Mapudungun	arn	mapu1245
Portuguese	por	port1283	Greek (Modern)	ell	mode1248
Indonesian	ind	stin1234	Armenian (Eastern)	hye	nucl1235
Turkish	tur	nucl1301	Hebrew (Modern)	heb	hebr1245
Vietnamese	vie	viet1252	Kannada	kan	nucl1305
Thai	tha	thai1261	Zulu	zul	zulu1248
Swahili	swh	swah1253	Igbo	ibo	nucl1417
Persian	pes	west2369	Ewe	ewe	ewee1241
Basque	eus	basq1248	Oromo (Harar)	hae	east2652
Yoruba	yor	yoru1245	Berber (Middle Atlas)	tzm	cent2194
Quechua (Imbabura)	qvi	imba1240	Maltese	mlt	malt1254
Arabic (Egyptian)	arz	egyp1253	Nepali	npi	nepa1254
Cantonese	yue	yuec1235	Sinhala	sin	sinh1246
Tamil	tam	tami1289	Lao	lao	laoo1244
Yucatec	yua	yuca1254	Ukrainian	ukr	ukra1253
Hausa	hau	haus1257	Cherokee	chr	cher1273
Amharic	amh	amha1245	Fijian	fij	fiji1243
Somali	som	soma1255	Aymara (Central)	ayr	cent2142
Wolof	wol	nucl1347	Nahuatl (Tetelcingo)	nhg	tete1251
Tagalog	tgl	taga1270	Pitjantjatjara	pjt	pitj1243
Javanese	jav	java1254	Dyirbal	dbl	dyir1250

A.3 External references and alternative explanations

WALS and Glottolog distinguish structural resemblance from genealogical classification (Dryer & Haspelmath, 2013; Hammarström et al., 2026). WALS v2020.4 distance is the proportion of jointly observed features on which two languages differ, excluding missing observations pairwise. Glottolog 5.3 distance measures genealogical separation by counting classification links between languages. Grambank provides a second description of grammatical structure (Skirgård et al., 2023a); its v1.0.3 release supports comparison on 52 shared languages (Skirgård et al., 2023b). These external linguistic descriptions test which documented language properties the model relations reflect.

The joint analysis of language distances asks whether WALS contributes beyond ancestry, geography, writing system, and material familiarity. It includes Glottolog path length and geographic distance. Writing-system predictors compare script distributions, primary-script mismatch, and character-trigram distance. Model-input predictors compare subword distributions and differences in tokenizer fertility (segmentation length), unknown-token use, and negative log likelihood of the aligned materials. Shared WALS coverage accounts for unequal documentation. Material likelihood serves as a proxy for familiarity; the actual training exposure of these items is unknown. Inference uses 10,000

Table S2: Models. Full names identify the released models; the short names are used from Section 2.1 onward. Model and family documentation is distinguished in Appendix A.2. Parameter counts are the stored weights of each model to two decimals (GPT-2-ZH from its configuration, with tied embeddings); architecture values are read from each configuration, with Heads giving query/key–value heads and Feed-forward the configured intermediate width. The subject and later-token bridge edits use the same zero-based block; the final block only records states. GPT-2-ZH was not intervened on.

Model	Short name	Family	Params
Llama-3-8B (Grattafiori et al., 2024)	Llama-8B	Llama 3	8.03B
DeepSeek-R1-Distill-Llama-8B (Guo et al., 2025)	DS-Llama	Llama 3	8.03B
Llama-3-Chinese-8B (Cui et al., 2023; Cui & Yao, 2024)	Llama-ZH	Llama 3	8.03B
Mistral-Nemo-Base-2407	Mistral	Mistral	12.25B
Qwen2.5-7B (Yang et al., 2024)	Qwen-7B	Qwen2.5	7.62B
Qwen2.5-7B-Instruct (Yang et al., 2024)	Qwen-7B-It	Qwen2.5	7.62B
Qwen2.5-14B (Yang et al., 2024)	Qwen-14B	Qwen2.5	14.77B
Gemma-2-9B (Gemma Team, 2024)	Gemma-9B	Gemma 2	9.24B
GPT-2-large-Chinese (Radford et al., 2019; Zhao et al., 2019)	GPT-2-ZH	GPT-2	0.74B

Model	Layers	Heads (Q/KV)	Hidden	Feed-forward	Edit block
Llama-8B	32	32/8	4096	14336	16
DS-Llama	32	32/8	4096	14336	16
Llama-ZH	32	32/8	4096	14336	16
Mistral	40	32/8	5120	14336	20
Qwen-7B	28	28/4	3584	18944	14
Qwen-7B-It	28	28/4	3584	18944	14
Qwen-14B	48	40/8	5120	13824	24
Gemma-9B	42	16/8	3584	14336	21
GPT-2-ZH	36	20/20	1280	5120	—

Table S3: Identity of the model weights behind the sentence representations and the confined edits. Blocks gives the block count implied by each model's stored representation of the 320 sentence occurrences, the count in the analysis configuration, and the count in Table S2; Width is the number of columns of the Subspace representation, equal to the sum of selected units over blocks. The stored representations match the checksums recorded when they were computed. Selected units gives the number selected for language relations at the edit block, over the units in that block; Llama-8B and GPT-2-ZH received no confined edits. No checksum of the weights was recorded when the representations were computed, so the identity rests on each model's configuration, block count and representation width; full names are in Table S2.

Model	Blocks	Width	Selected units at edit block
Llama-8B	32 / 32 / 32	3779	—
Llama-ZH	32 / 32 / 32	3765	1,440 of 14,336 (block 16)
DS-Llama	32 / 32 / 32	3736	813 of 14,336 (block 16)
Qwen-7B	28 / 28 / 28	3584	398 of 18,944 (block 14)
Qwen-7B-It	28 / 28 / 28	3584	369 of 18,944 (block 14)
Qwen-14B	48 / 48 / 48	6144	190 of 13,824 (block 24)
Mistral	40 / 40 / 40	4608	328 of 14,336 (block 20)
Gemma-9B	42 / 42 / 42	5338	59 of 14,336 (block 21)
GPT-2-ZH	36 / 36 / 36	4608	—

language-label permutations and Benjamini–Hochberg correction within the specified comparison families (Benjamini & Hochberg, 1995).

A.4 Language composition changes the reference comparison

The composition of the language sample determines which linguistic contrasts are tested. The 60 languages span 24 families, so any two languages differ jointly in ancestry, geographic distribution,

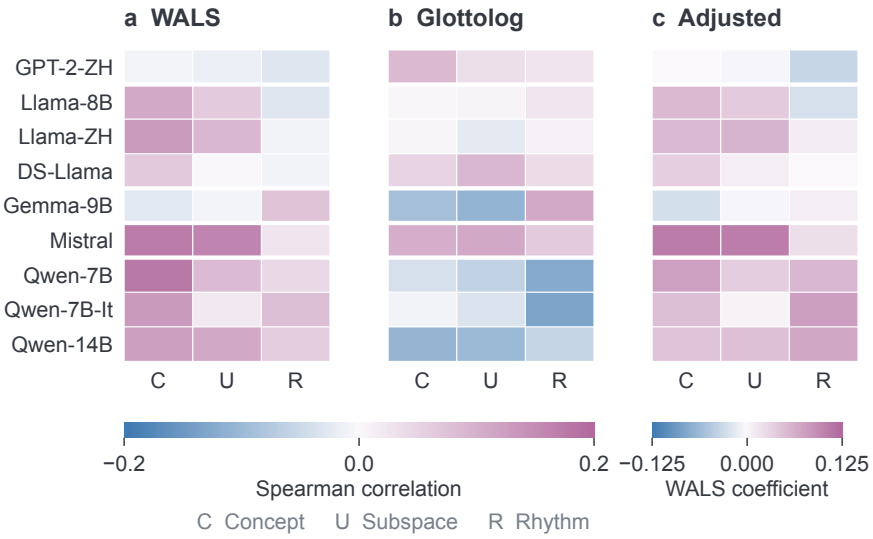

Figure S1: **Model agreement and correspondence with linguistic resources are distinct comparisons.** WALS and Glottolog correlations show every model and representation measure; the controlled WALS panel uses its separately labelled coefficient scale. The controlled panel applies BH correction to a family of 27 comparisons. Shared language nodes and related model families preclude treating the cells or two-language comparisons as independent replications.

writing systems, and documentation; correspondence with external references therefore reflects both the composition and the size of the sample.

Different representational properties become prominent at the family and language levels. Averaging distances across models and then, without weighting, across the languages of any two families gives one distance for every two of the 24 families, 276 in all. At this scale, Rhythm corresponds most strongly to Glottolog and Subspace most strongly to WALS, differing from the ordering at the language level. Broad family contrasts and differences between individual languages thus capture different aspects of model organization. Complete coefficients are available online.

A.5 Recurrent relations persist across meanings and language composition

Language relations derived from Concept define the most recurrent organization across the sampled model families. Its recurrence is .422, compared with within-model cross-fold reliability of .660: models preserve their own language ordering more closely than they agree with other families. Recurrence averages the correlations between models of two families and weights the ten family-to-family comparisons equally, so families with different numbers of models contribute equally.

Changing the meanings or removing the largest language family preserves the ordering of the three measures. All three exceed the random units in each of five folds. Different evaluation meanings lower Concept recurrence to .313, and removing all 13 Indo-European languages leaves it at .376. Meaning overlap strengthens recurrence, and Concept retains its advantage when the evaluation meanings differ. Fold ranges describe variation across overlapping discovery sets.

The Concept pattern extends across the sampled model families and languages. Every family's mean association with the other four is positive; Fisher-z averaging and omission of individual languages or families preserve that orientation. Recurrence thus extends across model families and remains positive after individual languages are omitted.

Cross-model recurrence and linguistic correspondence describe different aspects of model organization (Figure S1). After adjustment, the largest mean WALS coefficient is .0457 for Concept, with incremental $R^2 = .00314$. BH correction covers a family of 27 model-by-representation comparisons and a family of 513 coverage-eligible WALS-domain comparisons. On the common 52-language

subset, WALS and Grambank give different coefficients. Shared model organization therefore has a reproducible relational description beyond what these measured structural properties capture.

Unit selection affects cross-model recurrence and WALS correspondence differently. Before deactivation, selected units have lower WALS correlations than random units in all three representations, averaging five folds and weighting five model families equally. Full model, fold, resource, and clustering results are provided in the language-relations data of the anonymous repository.

A.6 Grammatical feature combinations and historical change

The historical comparisons address three questions: which grammatical properties coexist, how readily feature combinations change, and which ancestral states can be reconstructed. Historical proposals first predict which grammatical properties tend to coexist. We test whether model representations particularly reflect combinations supported by independent evolutionary analyses (Verkerk et al., 2026). One proposal predicts that languages marking dual number, distinguishing two referents, also mark plural number. Presence or absence of each property defines four combinations. On the 52-language overlap, we compare languages sharing versus differing in these combinations, accounting for genealogical, geographic, and background structural resemblance. The 60 externally supported proposals have support from both spatiophylogenetic and BayesTraits analyses; background and matched unsupported proposals test the specificity of their model associations.

The Llama 3 and Gemma 2 families assign more similar Rhythm representations to languages sharing feature combinations under the supported proposals, with associations surviving correction (Table S4). Gemma 2 also shows an association for matched unsupported proposals. The background comparison and the matched-proposal comparison give minimum $q = .09750$ and $.05700$ in separate 15-comparison families. Inference groups languages by family and macroarea; the matched comparison matches each of the 60 supported proposals with an unsupported one, 42 of them on the same linguistic domain.

A transition-based comparison tests a different historical prediction: feature combinations that are more readily connected by change should be closer in the model representation. Transition-rate estimates for 59 proposals define continuous-time Markov-chain commute-time distances between blocks of languages grouped by family and macroarea. These distances summarize relative movement between grammatical states after within-proposal normalization. After adjustment for static feature similarity and the other linguistic predictors, the 15 model-family-by-representation comparisons give minimum $q = .56282$ (Table S5).

Ancestral reconstruction evaluates the historical reference itself. For four proposals and 100 posterior trees per proposal, near-uniform root probabilities leave deep ancestral states unresolved. This calculation concerns the historical reference alone. The historical evidence separates present grammatical combinations, estimated transitions, and ancestral states; the observed Rhythm associations concern the first of these levels.

B Input changes and the recovery of language representations

B.1 Input transformations affect relation measures differently

The same input change can preserve overall activation organization while altering meaning relations or sequential response strength (Figure S2a). The four conditions are unaltered input, a non-identity random permutation of each four-unit sequence, replacement of the final two-unit phrase, and the fixed middle-unit swap $[1, 3, 2, 4]$. For example, “car auto turn bend” becomes “car auto dance move” after phrase replacement or “car turn auto bend” after swapping. Randomization changes Concept most strongly; Rhythm responds more to swapping than phrase replacement; Subspace remains comparatively stable. These distinct sensitivities show that the three measures report different aspects of language organization. Because the transformations jointly alter several linguistic properties, each measure reflects their combined consequences.

The model summaries average the predefined unit counts and layer tiers. The complete comparison across all conditions also evaluates changes between representation measures and is available online.

Table S4: Adjusted associations with grammatical configurations. Means average sign-reversed coefficients of shared configuration predicting rank-standardized model distance after linguistic adjustment. Positive values indicate smaller model distances between blocks sharing a configuration. Intervals are 95% percentile bootstrap intervals over configuration clusters. Each reference set has 15 upper-tail, BH-corrected tests permuting state labels among language-family × macroarea blocks within macroareas. Intervals describe variation across configurations; tests assess alignment under state relabeling. Table S5 tests specificity to externally supported proposals.

		Supported configurations		Matched configurations	
Model family	Representation	Mean [95% CI]	q	Mean [95% CI]	q
GPT-2	Concept	-0.021 [-0.066, 0.022]	0.9998	-0.033 [-0.087, 0.018]	1.0000
GPT-2	Subspace	-0.011 [-0.046, 0.023]	0.9998	0.006 [-0.029, 0.043]	1.0000
GPT-2	Rhythm	-0.057 [-0.118, -0.004]	0.9998	-0.017 [-0.061, 0.026]	1.0000
Llama 3	Concept	-0.102 [-0.181, -0.026]	0.9998	-0.142 [-0.223, -0.059]	1.0000
Llama 3	Subspace	-0.062 [-0.129, 0.008]	0.9998	-0.166 [-0.228, -0.103]	1.0000
Llama 3	Rhythm	0.068 [-0.007, 0.141]	0.0345	-0.089 [-0.196, 0.017]	1.0000
Gemma 2	Concept	0.005 [-0.069, 0.079]	0.9998	-0.128 [-0.221, -0.034]	1.0000
Gemma 2	Subspace	-0.060 [-0.135, 0.021]	0.9998	-0.146 [-0.231, -0.066]	1.0000
Gemma 2	Rhythm	0.089 [0.013, 0.163]	0.0030	0.206 [0.111, 0.308]	0.0015
Mistral	Concept	-0.094 [-0.178, -0.011]	0.9998	-0.146 [-0.241, -0.053]	1.0000
Mistral	Subspace	-0.115 [-0.187, -0.044]	0.9998	-0.176 [-0.260, -0.094]	1.0000
Mistral	Rhythm	0.034 [-0.030, 0.101]	0.9998	-0.048 [-0.104, 0.007]	1.0000
Qwen2.5	Concept	-0.074 [-0.150, 0.003]	0.9998	-0.114 [-0.190, -0.040]	1.0000
Qwen2.5	Subspace	0.001 [-0.086, 0.088]	0.9998	-0.083 [-0.165, -0.006]	1.0000
Qwen2.5	Rhythm	0.032 [-0.021, 0.085]	0.9998	0.004 [-0.047, 0.056]	1.0000

B.2 Deactivation and representation alterations address distinct relations

Candidate units capture two kinds of response: periodic variation across aligned input positions (frequency candidates) and differences among six predefined meaning categories (cluster candidates). Cross-language selections require inclusion in at least ten languages for deactivation and two for reconstruction; language-indexed selections contain the remaining selected units for each language. Their two intersections complete six overlapping groups, supplemented by a random reference. These selections concern unit responses across languages; the recurrence analysis separately compares relations across models. Selection details accompany the figure data in the anonymous repository.

Deactivation reveals heterogeneous contributions to the measured relations: candidate-minus-random differences are positive, near zero, or opposite in sign across models (Figure S2b). The random reference targets the union of all six candidate sets, matching layers and approximately activation strength; the available pool limits its realized size. It therefore differs in size from individual candidates. The comparison describes sensitivity to these particular unit sets, while unequal sizes preclude a size-controlled ranking of their importance. Agreement with WALS is assessed separately from the internal relation changes.

Projection removal, profile flattening and circular shifts of sequence positions test the contribution of selected components to external linguistic correspondence. Concept and Subspace use projection removal. The two Rhythm operations respectively replace each selected unit's position values by their mean or circularly shift them by half a sequence; both use a random-unit flattening reference. The contrast is $\Delta\rho_{\text{targeted alteration}} - \Delta\rho_{\text{random reference}}$, and all eight intervals include zero (Table S6). Internal sensitivity and external linguistic correspondence therefore provide different evidence about the contribution of the edited component.

B.3 Language of origin and unit overlap shape reconstruction

Reconstruction tests how well source activation values recover the recipient's original representation. It crosses three layer tiers, seven candidate sets, and three recipient input conditions. Source values replace selected recipient values at the same layer with mixing coefficient one. The endpoint is

Table S5: Historical specificity and transition-distance comparisons. Background Δ is the supported-versus-unsupported coefficient for configuration association, adjusted for domain and coverage; Matched Δ averages supported-minus-matched differences. Positive Δ favors supported proposals. CTMC entries summarize adjusted regression coefficients relating rank-standardized model distance to rank-standardized transition distance; positive coefficients indicate that the distances increase together. Each q column applies BH correction to 15 upper-tail tests. Posterior intervals describe variation across CTMC draws, whereas macroarea-restricted QAP tests correspondence under reference relabeling.

		Background		Matched		CTMC reference	
Model family	Representation	Δ	q	Δ	q	Median [95% interval]	q
GPT-2	Concept	-0.001	0.763	0.012	0.441	0.099 [0.002, 0.185]	0.772
GPT-2	Subspace	-0.019	0.874	-0.017	0.896	-0.124 [-0.201, -0.053]	0.903
GPT-2	Rhythm	0.029	0.593	-0.040	0.919	-0.491 [-0.591, -0.389]	0.952
Llama 3	Concept	0.027	0.728	0.040	0.324	0.504 [0.311, 0.644]	0.563
Llama 3	Subspace	0.086	0.199	0.104	0.058	0.325 [0.145, 0.470]	0.563
Llama 3	Rhythm	0.137	0.097	0.157	0.057	-0.116 [-0.291, 0.035]	0.772
Gemma 2	Concept	0.098	0.237	0.133	0.133	-0.127 [-0.335, 0.096]	0.772
Gemma 2	Subspace	0.024	0.728	0.087	0.204	0.014 [-0.129, 0.195]	0.772
Gemma 2	Rhythm	-0.086	0.925	-0.117	0.978	0.486 [0.322, 0.663]	0.759
Mistral	Concept	-0.038	0.856	0.053	0.324	0.450 [0.126, 0.671]	0.563
Mistral	Subspace	-0.024	0.839	0.061	0.315	-0.031 [-0.314, 0.161]	0.759
Mistral	Rhythm	-0.001	0.763	0.082	0.133	-0.283 [-0.409, -0.139]	0.772
Qwen2.5	Concept	0.015	0.728	0.040	0.324	0.240 [0.001, 0.466]	0.563
Qwen2.5	Subspace	0.084	0.237	0.085	0.204	-0.402 [-0.582, -0.210]	0.772
Qwen2.5	Rhythm	-0.006	0.763	0.028	0.324	-0.011 [-0.113, 0.073]	0.772

Table S6: External-alignment changes after representation alterations relative to the specified random references. Differences are Fisher-z averages across nine models; 95% percentile intervals resample models (2,000 draws), keeping languages fixed. Dyads gives the number of two-language comparisons in each alignment. All intervals include zero.

Operation	Reference	Difference	95% interval	Dyads
Concept	Glottolog	-0.001126	[-0.002960, 0.000470]	1770
Concept	WALS	0.000290	[-0.002815, 0.003248]	1769
Sequence order	Glottolog	0.000645	[-0.000007, 0.001454]	1770
Sequence order	WALS	-0.000105	[-0.000271, 0.000055]	1769
Rhythm profile	Glottolog	0.000324	[-0.000291, 0.000916]	1770
Rhythm profile	WALS	0.000048	[-0.000236, 0.000318]	1769
Subspace	Glottolog	0.000342	[-0.001532, 0.002082]	1770
Subspace	WALS	-0.001602	[-0.004813, 0.001344]	1769

algebraic similarity to the recipient's own full reference at that layer; subsequent network activity and behavior are outside this measurement.

For baseline, reconstructed, and full-reference similarities s_0, s_{edit}, and s_{full}, the unnormalized similarity change (raw lift) is $s_{\text{edit}} - s_0$; eligible gap closure is $(s_{\text{edit}} - s_0)/(s_{\text{full}} - s_0)$. Mean change covers more observations. For Rhythm gap closure, denominator eligibility excludes 25,907 of 34,020 same-language ratios and 284,977 of 374,220 cross-language ratios. The two summaries quantify different reference comparisons on different sets of observations.

Source language consistently conditions local recovery: each model has positive mean change for same-language sources and negative mean change for cross-language sources. Concept shows the strongest contrast ($+.105$ versus $-.139$), with the other representation measures and eligible-gap means oriented alike. Same-language sources share more of the values underlying the recipient's

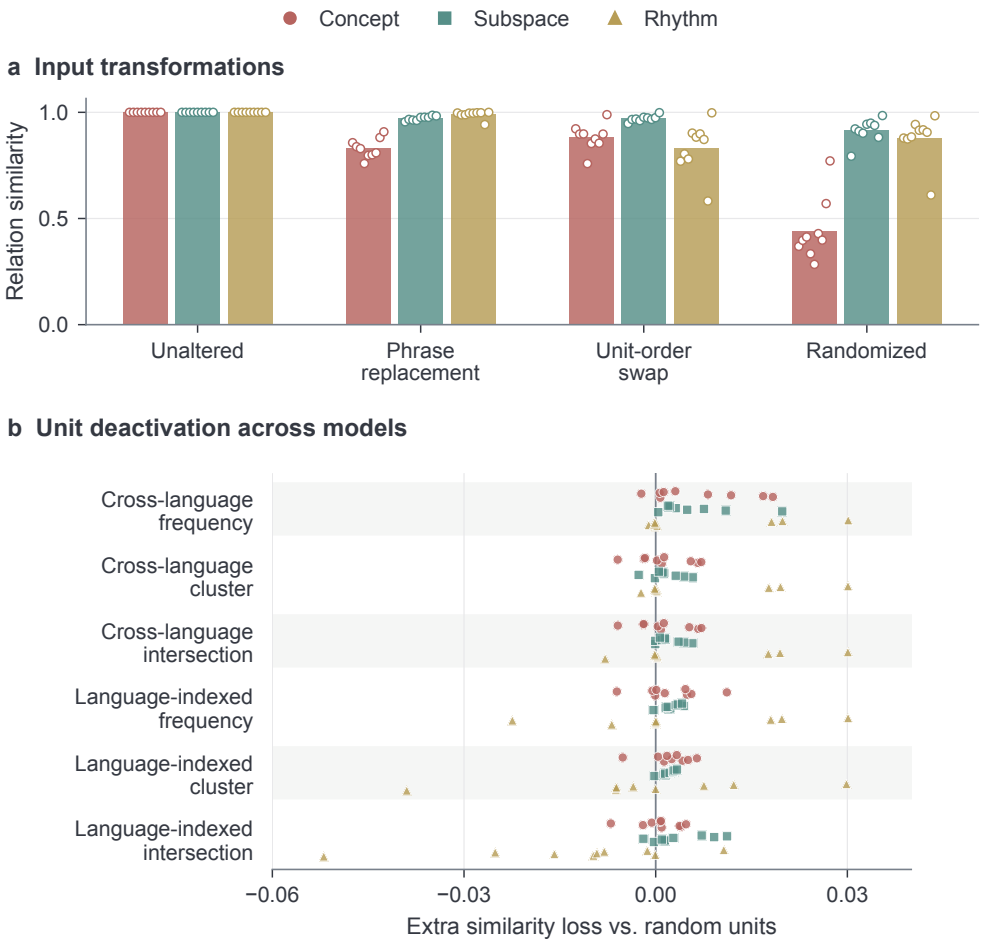

Figure S2: **Input changes and unit deactivation reveal different forms of sensitivity.** **a**, Four input transformations across three relation measures. **b**, Internal relation loss after candidate-unit deactivation, subtracting a random deactivation matched by model, language, and layer tier. The response is change in the fixed representation measure; WALS correspondence is evaluated separately. Points describe model variation over repeated language/tier conditions.

own reference. Shared language relations therefore coexist with language-dependent recovery of a particular representation.

Candidate selection explains substantial variation within this source-language contrast (Figure S3). Sets with minimal overlap with the units used to measure the representation often produce changes near zero; other sets produce changes of either sign. Cross-language replacement has positive responses in all three measures, and its Concept median is zero. The negative average combines heterogeneous conditions, including successful cross-language recovery and many unchanged responses.

The online reconstruction data provide all 408,240 source rows (34,020 same-language and 374,220 cross-language) and candidate-by-model summaries. Additional absolute-similarity and distance-binned Concept summaries cover the random-input subset, whereas the mean-change comparison includes all recipient conditions. Distance-bin intervals are empirical interquartile ranges; inclusive edges can share boundary observations. These summaries characterize different reference comparisons, with their original observations retained.

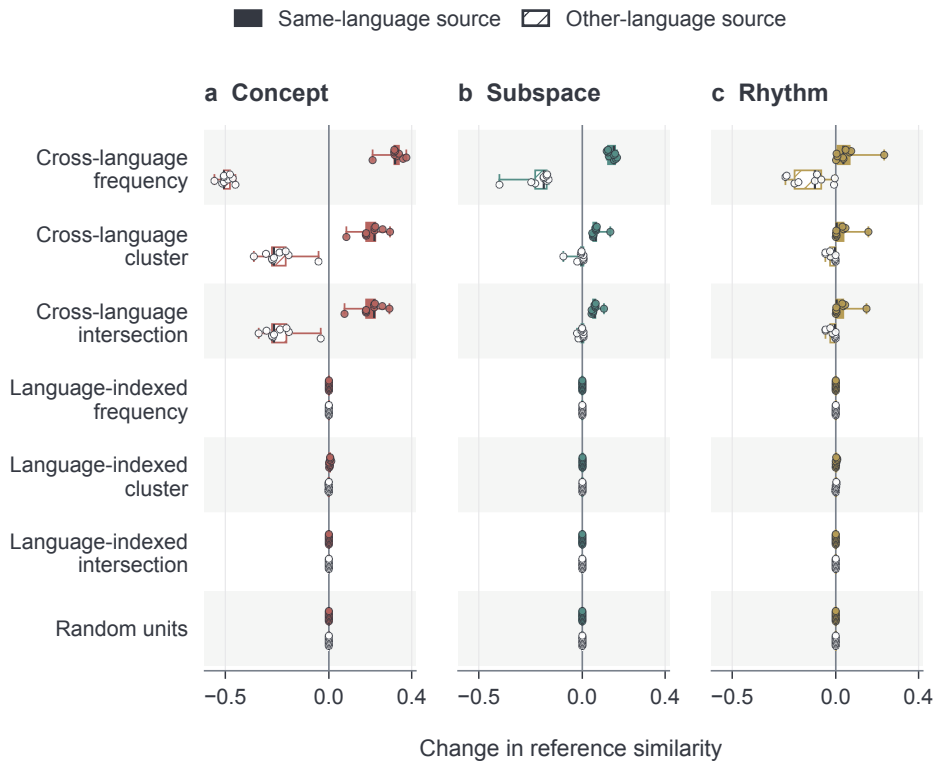

Figure S3: **Candidate sets retain substantial variation within the source-language comparison.** Raw-lift summaries are computed jointly for model, candidate set, and source-language condition (same or other language) across replacement conditions. Boxes describe empirical quantiles across the nine related models. Zero and opposite-sign responses remain visible. Six overlapping candidate sets are defined by unit responses across aligned positions and meaning categories; the seventh uses randomly chosen units.

B.4 Broad edits have task-dependent behavioral effects

Broad representation edits affect question answering differently from paraphrase and inference (Table S7). Positive control-corrected drops indicate additional impairment: differences are positive for MLQA and XQuAD, negative for XNLI, and vary across operations for Belebele and PAWS-X. Coverage ranges from 7 to 45 languages. The behavioral consequences depend on the task being scored, motivating a direct test of grammatical influence on agreement.

The four operations alter Concept relations, Subspace alignment, Rhythm profiles, and sequence order. All 180 model/task/operation comparisons, their language coverage, and control-corrected components are available online. Tasks differ in sampled languages and scored decisions, so their effects quantify distinct behavioral sensitivities.

C Number intervention and the conditions of transfer

C.1 Materials, intervention sites, and functional questions

The number experiments ask whether the units that carry the recurrent relations concentrate grammatical influence, and how represented grammatical information influences agreement across forms and contexts. English tests exclude a form of number marking from estimation; Hebrew comparisons test cross-language transfer and its dependence on strength and surrounding number information. Availability asks whether a direction fitted on some forms of number marking also distinguishes number in an excluded form. Selective intervention asks whether changing the subject along that

Table S7: Mean control-corrected accuracy changes across the available models, in percentage points. Positive values mean a larger performance loss after the targeted edit. Each task mean describes its own language sample. The tasks are Belebele (Bandarkar et al., 2024), MLQA (Lewis et al., 2020), PAWS-X (Yang et al., 2019), XNLI (Conneau et al., 2018) and XQuAD (Artetxe et al., 2020).

Task	Languages	Concept	Subspace	Rhythm profile	Sequence order
belebele	45	-0.015	+0.040	-0.026	+0.057
mlqa	7	+0.596	+0.634	+0.162	+0.343
pawsx	7	-0.255	-0.509	-0.107	+0.095
xnli	13	-0.298	-0.298	-0.467	-0.539
xquad	11	+0.496	+0.572	+0.020	+0.094

direction affects agreement more than equal-sized changes in another direction or at another position. Compensation asks whether a later edit restores agreement specifically after subject disruption. Contextual modulation asks how surrounding noun number alters these responses.

English number is expressed through regular nouns (R), irregular nouns (I), and invariant-head demonstratives (D), with 12 subject-lexeme clusters per group and 36 in total. Their factorial combinations yield 24 contexts per cluster, 864 in all. Literal examples are "Yesterday the pilot near the guards clearly appears", "Yesterday the child near the parents clearly appears", and "Yesterday this sheep near the goats clearly appears", with the alternative target "appear". All items share this carrier frame, and every effect is a difference between edited and reference conditions on the same frame, so frame-specific preferences are common to the compared conditions. Singular/plural head forms include pilot/pilots, child/children, and this/these sheep. The temporal carrier is fixed, and the scored contrast is agreement between the subject's number and the two verb forms.

The agreement margin is the complete target-word log probability of the agreeing form minus that of the disagreeing form. The fitted direction is oriented plural minus singular. Subject number s ($+1$ plural, -1 singular) determines the offset $h' = h - s\alpha v$ at the head-final state, shifting it against the subject's number. The intervention applies the same number-dependent offset to each state; its size is fixed across items. An opposite edit at a later token in the same block tests compensation; the last block only records states. For Qwen-7B, the subject, off-position and later "bridge" tokens are pilot/child/sheep, Yesterday and clearly (indices 2, 0 and 6), edited at block 14. Subject and bridge occupy different positions within the same layer.

C.2 Eligibility and transfer of readable number information

Eligibility requires an existing number-agreement preference and sensitivity to an intervening noun. Models must have positive mean agreement in all 18 cue-stratum/attraction/subject-number cells and a positive classic singular-subject attraction contrast, $M_{\mathrm{matched,SG}} - M_{\mathrm{mismatched,SG}}$. The latter uses 99,999 lexeme-level sign randomizations and a plus-one probability at .05. At block $\lfloor L/2 \rfloor$, number directions fitted on two marking groups are evaluated on the third. This availability test uses 99,999 cluster-sign draws at .01 and requires positive effects in every held-out group (Table S8).

Of the initial Llama-8B evaluation's four endpoints, attenuation, position selectivity and compensation were supported; attraction-conditioned modulation gave DiD .0163 (one-sided $p = .07999$). The eight-model extension was specified before its outcomes were inspected. Seven models proceeded to intervention; GPT-2-ZH did not meet the baseline agreement criterion and was not intervened on.

C.3 Earlier full-fit intervention and its comparison conditions

The earlier intervention fitted its number direction and comparison directions on all 36 clusters; the held-out extension below fits them only on the other marking groups. In the earlier design, edit strength was half the median plural-minus-singular displacement along the number direction. Its 29 conditions comprised no-write sham, subject edit, the same edit off position, eight equal-length orthogonal subject edits, bridge-only and joint edits, eight orthogonal bridge edits, and eight joint subject-plus-orthogonal-bridge edits. These comparisons separate the number direction's influence from the location and general benefit of an edit.

Table S8: English baseline attraction and availability of number information. Attraction is the matching-minus-mismatching intervening-noun margin for singular subjects. Number availability (T) is the mean standardized number projection across held-out cue types; it is distinct from intervention transfer across languages. Each regular, irregular and demonstrative stratum contains 12 lexical clusters. GPT-2-ZH did not meet the requirement of positive agreement in all 18 baseline conditions and was not intervened on.

Model	Baseline attraction	p	Number availability (T)	p
Qwen-7B	0.8334	10^{-5}	2.8902	10^{-5}
Mistral	1.2370	10^{-5}	2.9367	10^{-5}
Gemma-9B	2.0973	10^{-5}	2.6712	10^{-5}
Qwen-7B-It	0.9720	10^{-5}	2.9353	10^{-5}
Qwen-14B	0.6378	10^{-5}	3.5398	10^{-5}
Llama-ZH	0.6270	10^{-5}	2.7635	10^{-5}
DS-Llama	1.0397	10^{-5}	3.0089	10^{-5}
GPT-2-ZH	0.1729	10^{-5}	n/a	n/a

Table S9: Earlier English intervention effects using operators fitted on all 36 clusters. Holm correction covers the three endpoints within each model. These results establish the original fitting regime, separately from the held-out evaluation.

Model	A	p_H	S	p_H	C	p_H
Qwen-7B	0.23045	3×10^{-5}	0.22913	3×10^{-5}	0.02505	3×10^{-5}
Mistral	0.26561	3×10^{-5}	0.25281	3×10^{-5}	0.00823	3×10^{-4}
Gemma-9B	0.52032	3×10^{-5}	0.46248	3×10^{-5}	0.02567	3×10^{-5}
Qwen-7B-It	0.25760	3×10^{-5}	0.25561	3×10^{-5}	0.03424	3×10^{-5}
Qwen-14B	0.11470	3×10^{-5}	0.11029	3×10^{-5}	0.00926	3×10^{-5}
Llama-ZH	0.24062	3×10^{-5}	0.22760	3×10^{-5}	0.02680	3×10^{-5}
DS-Llama	0.38111	3×10^{-5}	0.36429	3×10^{-5}	0.10013	3×10^{-5}

Attenuation measures the sham-minus-subject agreement margin. Selectivity takes the smaller of the advantages over alternative-direction and alternative-position edits. Compensation takes the smaller of the fitted interaction and its excess over orthogonal bridge interactions. One-sided cluster sign flips use 99,999 draws, followed by Holm adjustment (Holm, 1979) across the three endpoints within a model (Table S9). These intervention results use all clusters for fitting and evaluation; held-out causal effects are assessed separately below.

C.4 Intervention generalization to held-out cue types

The held-out extension asks whether grammatical influence extends to a form of number marking absent from estimation. Each fold fits 24 clusters from two marking groups and evaluates the remaining 12. Evaluation clusters contribute nothing to the direction, coordinate variance, edit strength, or either bank of comparison directions. The same factorial contexts and edits are evaluated in every fold. Because training sets overlap, the pooled 36-cluster mean describes the cross-fitted effects; inference uses the individual held-out groups.

For training differences $\delta_i = h_{i,\mathrm{PL}} - h_{i,\mathrm{SG}}$, coordinate means μ_j, unbiased variances s_j^2, and their coordinate average $\bar{s^2}$, the estimator is

$$q_j = .5s_j^2 + .5\bar{s^2} + 10^{-12}, \qquad \widetilde{v}_j = \mu_j/q_j, \qquad v = \widetilde{v}/\|\widetilde{v}\|_2, \qquad \alpha = .5\,\mathrm{median}_i(\delta_i^\top v).$$

Each eight-direction bank begins with independent $\sqrt{q} \odot z$, $z \sim \mathcal{N}(0, I)$, followed by orthogonalization against v and previously accepted directions, normalization, and a fixed sign convention. Both banks are fitted from training clusters. These alternatives reflect the coordinate variation in the training data while changing the direction of the intervention. Complete materials, seeds, and fitting settings are supplied in the anonymous repository.

Table S10: Held-out English effects and all-cue inference. Means summarize the three folds. For each endpoint, the largest fold probability tests support across all three cue strata; Holm correction covers A/S/C within each model. Compensation meets this criterion in six of the seven models.

Model	A	p_H	S	p_H	C	p_H
Qwen-7B	0.196778	0.000732	0.195783	0.000732	0.018357	0.025146
Llama-ZH	0.213270	0.000732	0.200485	0.000732	0.025568	0.004639
DS-Llama	0.331812	0.000732	0.315871	0.000732	0.090306	0.000732
Gemma-9B	0.562630	0.000732	0.504453	0.000732	0.027757	0.001709
Mistral	0.257637	0.000732	0.243746	0.000732	0.010027	0.092773
Qwen-7B-It	0.211558	0.000732	0.210227	0.000732	0.024453	0.031250
Qwen-14B	0.103561	0.000732	0.099908	0.000977	0.007469	0.001953

Let m_0, m_s, m_o, m_k denote no-write sham, subject edit, off-position edit and orthogonal subject edit k; brackets average the prescribed contexts within lexical cluster i. The general endpoints are

$$A_i = \langle m_0 - m_s \rangle_i, \tag{6}$$

$$S_i = \min \left\{ \left\langle \frac{1}{8} \sum_{k=1}^{8} m_k - m_s \right\rangle_i , \langle m_o - m_s \rangle_i \right\}. \tag{7}$$

For singular subjects with mismatched intervening-noun number, the bridge interaction is

$$R_{v,i} = \langle (m_{s+b} - m_s) - (m_b - m_0) \rangle_{i,\mathrm{SG,mis}}, \tag{8}$$

$$R_{\mathrm{comparison},i} = \left\langle \frac{1}{8} \sum_{k=1}^{8} \{ (m_{s+b_k} - m_s) - (m_{b_k} - m_0) \} \right\rangle_{i,\mathrm{SG,mis}}, \tag{9}$$

$$C_i = \min\{R_{v,i}, R_{v,i} - R_{\mathrm{comparison},i}\}. \tag{10}$$

The minimum is taken before averaging clusters. Subtracting the bridge-only effect asks whether the later edit helps more after subject disruption than it does on its own. Thus C isolates the additional benefit of the later edit that depends on the earlier disruption.

All seven models generalize attenuation and selectivity across the three forms of number marking; six generalize compensation (Table S10). Each fold uses all $2^{12} = 4,096$ one-sided cluster sign flips. The maximum fold probability and positive means in every fold test support across all marking groups, followed by Holm correction for A/S/C within each model. Sign exchangeability is assumed at the lexical-cluster level.

The individual lexical results explain the difference between disruption and compensation. Attenuation is positive in every model-by-cluster observation, and selectivity has just one negative cluster, for Qwen-14B regular nouns. Compensation varies more: Mistral's irregular-noun mean is positive ($.004442$), with seven supporting and five opposing clusters giving $p = .092773$. The all-group criterion therefore separates a positive average from consistent cluster-level support. All fold endpoints, including the secondary attraction contrast and fitting comparisons, are available online; the models repeatedly evaluate the same 36 lexical clusters.

C.5 Confining the edit to the units selected for language relations

The confined edit asks whether grammatical influence is concentrated in the output space of units selected for language relations. The units are the intermediate feed-forward neurons selected in at least ten of the 60 languages by the frequency or meaning-category criterion of Appendix B.2, taken at each intervention model's edit block: 1,440 units in Llama-ZH, 813 in DS-Llama, 59 in Gemma-9B, 328 in Mistral, 398 in Qwen-7B, 369 in Qwen-7B-It and 190 in Qwen-14B, out of 13,824–18,944 units per block (Table S3). Each unit writes into the residual stream along one column of the block's output projection. The confined edit projects the fitted number direction onto the space spanned by those columns and renormalizes it, so the edit keeps its fitted strength and length and changes only its direction; the cosine between confined and unconfined directions lies

between .12 and .60 across models and folds. For each model and fold, twenty sets of randomly chosen units of the same number, at the same block, and with the same distribution of mean absolute activation over the training clusters' prompts (matched in ten activation bands) give the reference distribution; each set defines its own confined direction in the same way, and no set overlaps the selected units. The folds, the fitted direction and strength, the comparison directions, the 4,096 sign flips, the all-form criterion and the Holm correction are those of Appendix C.4, applied unchanged to every edit; the unconfined edit recomputed under the same conditions reproduces Table S10 to within 3×10^{-6}. For Qwen-7B, the confined direction replaces the English direction in the Hebrew transfer of Appendix C.10 at the common lower strength, with the Hebrew-derived edit, the baseline and the 415 evaluation configurations unchanged. The twenty random unit sets are those drawn for the model's first fold, since the Hebrew comparison is not divided into folds. All scoring used single precision; the scorer reproduces the released scorer to within 7×10^{-5} log-preference units on every model, and the compensation endpoint, whose scale is a few thousandths, was recomputed on the CPU for the confined and unconfined edits of every model and agrees to within 5×10^{-5}.

The comparison and its reading were fixed before any result was computed. A confined edit exceeding all twenty random sets in a model, while still weakening agreement across all three forms, would support concentrated grammatical influence through those units; a confined edit below the median and below eighteen or more sets is exceeded by random units; the remaining cases lie within the random range. With twenty sets, the smallest attainable share of sets at or above the confined edit is one in twenty-one. Because set sizes differ by more than an order of magnitude across models, no statistic pools the models, and Gemma-9B's 59 units give it the least power, as recorded in advance.

The confined edit retains between 3% and 41% of attenuation and selectivity and at most 24% of compensation, and it still weakens agreement across all three forms of marking in five models (Table S11). In no model does it exceed all twenty random sets. For attenuation and selectivity it lies within the random range in five models and below eighteen or more of the twenty sets in Llama-ZH and DS-Llama; for compensation it lies within the range in four models and below eighteen or more sets in DS-Llama, Gemma-9B and Qwen-14B, and the confined compensation is negative in Gemma-9B, Mistral and Qwen-14B. In Qwen-7B, the confined edit orders the 415 Hebrew configurations like the Hebrew-derived edit ($\rho = .847$ for attenuation and .880 for selectivity; unconfined .924 and .917; random sets .24–.94 across the two endpoints, with means .64 and .60), while its mean effect is .50 and .47 of the Hebrew-derived effect (unconfined .87 and .82; random sets .35 and .32 on average). Edits confined to selected units retain reduced grammatical effects, within or below the range for randomly chosen units. In Hebrew, different unit sets preserve similar sentence rankings while changing effect size. Fractions are ratios of confined to unconfined effects.

C.6 Compensation and contextual modulation can diverge

Mistral distinguishes restoration after two edits from compensation for the first edit (Figure S4a). In irregular-noun materials, both joint recovery and bridge-only recovery are about .31 log-preference units. Their difference defines the additional interaction. Because the alternative-direction interaction is negative, the fitted number interaction determines compensation in every cluster. The later edit benefits agreement largely on its own, which is what the cross-form compensation criterion separates from recovery dependent on the earlier disruption.

Gemma-9B reveals that averaging can conceal opposing contextual responses (Figure S4b). A mismatching intervening noun amplifies the subject edit for regular and irregular forms but reduces its effect in all twelve demonstrative clusters, whose mean difference is $-.116400$ log-preference units. The pooled average is near zero. Both the fitted direction and the evaluated materials change across the three groups.

C.7 Earlier attraction-conditioned coupling

The earlier coupling endpoint divided the mean singular-subject mismatch-minus-match attenuation by mean general attenuation. A cluster bootstrap resampled both quantities together in 99,999 draws. The prespecified $[-.10, .10]$ equivalence region defined modulation smaller than one tenth of attenuation; an interval entirely inside that region met the equivalence criterion. Positive coupling required a positive increment and one-sided sign-flip $p \leq .05$.

Table S11: The number edit confined to the units selected for language relations. Units is the number of selected units at the model's edit block, and the cosine compares the confined with the unconfined direction across the three folds. Confined and Unconfined give pooled held-out effects in log-preference units; Fraction is their ratio. Twenty sets of randomly chosen units matched in number, layer and activation strength give the reference distribution, summarized by its mean and SD and by the number of sets at or above the confined value (eighteen or more marks an edit that random units exceed). The sign-test probability is the Holm-corrected maximum fold probability of the confined edit under the all-form criterion of Appendix C.4. The unconfined values reproduce Table S10 to within 3×10^{-6}. The Hebrew rows give, for Qwen-7B at the common lower strength, the Spearman correlation of sentence effects with the Hebrew-derived edit and the mean effect relative to the Hebrew-derived effect.

Model	Effect	Units	Cosine	Confined	Unconfined	Fraction	Random mean ± SD	Sets ≥ confined	Sign test p
Llama-ZH	A	1440	0.57–0.60	0.036731	0.213270	0.172	0.0596 ± 0.0134	18	0.0007
	S			0.031079	0.200485	0.155	0.0518 ± 0.0125	19	0.0007
	C			0.001792	0.025569	0.070	0.0027 ± 0.0015	14	0.1235
DS-Llama	A	813	0.44–0.50	0.055263	0.331811	0.167	0.0978 ± 0.0269	19	0.0007
	S			0.050462	0.315868	0.160	0.0907 ± 0.0253	19	0.0007
	C			0.002182	0.090303	0.024	0.0167 ± 0.0080	20	0.8169
Gemma-9B	A	59	0.12–0.13	0.017593	0.562629	0.031	0.0317 ± 0.0277	15	0.0608
	S			0.015457	0.504454	0.031	0.0266 ± 0.0266	13	0.3633
	C			−0.000441	0.027755	−0.016	0.0004 ± 0.0009	18	0.9995
Mistral	A	328	0.33–0.39	0.035395	0.257637	0.137	0.0881 ± 0.0661	16	0.0168
	S			0.024808	0.243745	0.102	0.0739 ± 0.0625	16	0.1748
	C			−0.001660	0.010027	−0.166	-0.0006 ± 0.0034	16	0.9995
Qwen-7B	A	398	0.43–0.46	0.080141	0.196778	0.407	0.0825 ± 0.0481	9	0.0015
	S			0.079886	0.195782	0.408	0.0808 ± 0.0481	9	0.0015
	C			0.002837	0.018357	0.155	0.0046 ± 0.0033	14	0.3613
Qwen-7B-It	A	369	0.42–0.45	0.084233	0.211557	0.398	0.0583 ± 0.0411	5	0.0007
	S			0.083719	0.210225	0.398	0.0565 ± 0.0412	5	0.0007
	C			0.005819	0.024456	0.238	0.0029 ± 0.0040	1	0.1729
Qwen-14B	A	190	0.17–0.19	0.009866	0.103561	0.095	0.0130 ± 0.0214	8	0.3413
	S			0.007744	0.099906	0.078	0.0107 ± 0.0215	8	0.5068
	C			−0.000135	0.007469	−0.018	0.0009 ± 0.0009	18	0.9507
Qwen-7B, Hebrew transfer at the common lower strength; twenty random sets of 398 units									
Qwen-7B	Ordering, A (ρ)	398		0.847	0.924		0.637 ± 0.198	5	
Qwen-7B	Ordering, S (ρ)	398		0.880	0.917		0.601 ± 0.223	4	
Qwen-7B	Relative effect, A	398		0.501	0.867		0.347 ± 0.384	5	
Qwen-7B	Relative effect, S	398		0.474	0.819		0.323 ± 0.366	5	

Table S12: Earlier intervention-conditioned attraction relative to attenuation. The prespecified equivalence region is [-0.10,0.10]; Gemma-9B lies within it. This inference concerns the earlier all-cluster ratio; Table S10 reports the held-out cue results.

Model	Increment	p	Ratio	90% interval
Qwen-7B	0.04484	5×10^{-5}	0.19459	[0.11724, 0.28578]
Mistral	0.03955	8.2×10^{-4}	0.14889	[0.07868, 0.21993]
Gemma-9B	0.00465	0.37583	0.00893	[−0.03517, 0.05364]
Qwen-7B-It	0.04075	0.00202	0.15818	[0.06964, 0.26103]
Qwen-14B	0.01920	2×10^{-5}	0.16741	[0.11004, 0.23568]
Llama-ZH	0.02564	0.00219	0.10655	[0.04971, 0.16570]
DS-Llama	0.07257	10^{-5}	0.19041	[0.13746, 0.24349]

Context amplifies grammatical disruption in six models under the earlier all-cluster fit, including after a seven-model Holm sensitivity correction (largest adjusted $p = .00606$). Gemma-9B has substantial unedited attraction, and its additional modulation relative to the overall edit effect meets the equivalence criterion: the ratio's 90% interval lies inside the prespecified $\pm.10$ region (Table S12). Baseline attraction and modulation of an intervention therefore measure different forms of contextual sensitivity. The opposing responses across Gemma-9B's held-out marking groups show that its earlier near-zero average combines condition-specific responses of opposite sign.

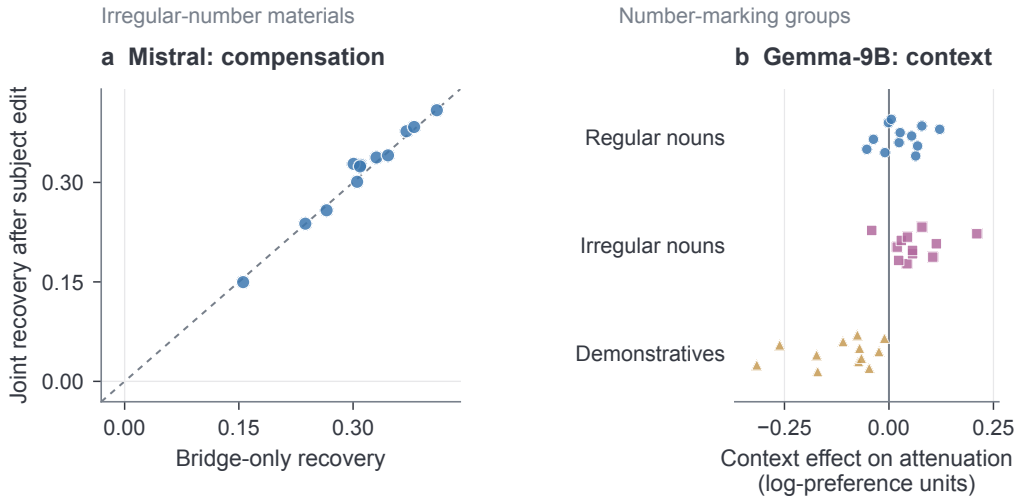

Figure S4: **Bridge-only gain and context-dependent attenuation decompose the compensation and modulation endpoints. a**, Mistral's twelve irregular-fold clusters compare joint recovery with bridge-only gain; their difference defines the fitted nonadditive interaction. **b**, Gemma-9B's singular-subject mismatch-minus-match attenuation across the three held-out cue folds, including the negative demonstrative responses. Both panels describe the components of the previously defined endpoints.

C.8 Hebrew-native intervention and construction sensitivity

Hebrew tests whether number intervention preserves its effect across sentence constructions. CLAMS provides simple agreement, VP coordination, and long-VP coordination (Mueller et al., 2020). Discovery uses 105 sentence combinations and evaluation uses 415 others, each with reciprocal singular/plural forms (830 items). Lexemes can recur between the two sets. All six construction/number-direction baseline cells favor agreement.

At Qwen-7B block 14, the normalized mean discovery plural-minus-singular difference defines the Hebrew direction; half its median projection defines strength 13.98469. Four conditions compare baseline, signed subject edit, an equal-length discovery-shaped orthogonal subject edit, and the same number edit off position. Reciprocal items are averaged into sentence-combination effects. Hebrew attenuation compares orthogonal with subject editing, whereas English uses the unedited reference; position selectivity compares off-position with subject editing. Original normalized effects divide each numerator by the mean absolute orthogonal margin.

The Hebrew-derived edit consistently changes agreement in the intended direction, while its size depends on the construction. Both endpoints are positive in 409 of 415 configurations, and 176/415 and 178/414 non-tied configurations exceed the predefined .10 threshold relative to the control margin. Simple-agreement sentences have larger mean effects than both VP-coordination constructions (Table S13). The two coordination constructions also differ in mean effect, showing that susceptibility varies with the particular sentence structure.

C.9 Transport of the English operation

The English-to-Hebrew comparison transfers the original English edit without fitting it to the Hebrew outcomes. Specified after the native result, it retains the English direction, sign, block, strength (8.54718), and comparison construction on the same Hebrew materials. Its oriented cosine with the native direction is .42241. No target-language rotation, strength calibration, layer search, or outcome-based item selection is used. This comparison uses the original English operator; the later cross-fitted operators belong to the separate English evaluation.

English-derived editing largely preserves which Hebrew constructions and sentences are most susceptible, with normalized rank correlations of approximately .88 for both endpoints. Effects are positive

Table S13: Original Hebrew comparisons at their separately fitted strengths. Native and Source denote the Hebrew- and English-derived edits, respectively. Attenuation and selectivity are divided by the mean absolute perpendicular-control margin over the two reciprocal number forms of each combination. Native effects exceed .10 in 176/415 and 178/414 non-tied configurations; transported effects in 53/415 and 55/415. Source/native ratios compare unequal strengths; their 90% percentile intervals use joint resampling of the 415 configurations (99,999 draws).

Construction	N	Native A	Native S	Source A	Source S
Simple agreement	53	0.56868	0.53293	0.16013	0.17315
Long VP coordination	80	0.12157	0.12173	0.04072	0.03713
VP coordination	282	0.09174	0.09375	0.02706	0.02623
All constructions	415	0.15840	0.15523	0.04669	0.04709

Original strong-effect criterion and source/native retention

Endpoint	Source mean	Positive	Above .10	Retention [90% interval]
Attenuation	0.04669	395/415	53/415	29.5% [26.9, 31.7]
Position selectivity	0.04709	398/415	55/415	30.3% [28.8, 31.9]

in 395/415 and 398/415 configurations, and 53/415 and 55/415 exceed the predefined .10 threshold. Their means are about 30% of the Hebrew reference, which uses a larger edit (Table S13). This agreement in response ordering, together with the reduced mean amplitude, distinguishes transferable sensitivity from the magnitude assigned by a particular strength and reference.

C.10 Edit strength changes the size of the cross-language difference

Common edit strengths test how much the source-language contrast depends on the size of the intervention. Both operations are evaluated at Euclidean norms $L = 8.547$ and $H = 13.985$, retaining their original direction estimators and perpendicular comparison directions. English uses variance shrinkage; Hebrew uses the mean plural-minus-singular difference. Equal lengths standardize the amount of state change, while the grammatical information affected can still depend on the estimated direction.

For each sentence combination, reciprocal number items are first averaged. With baseline, subject, orthogonal and off-position margins m_0, m_s, m_q, m_o, define

$$D = m_0 - m_s, \quad Q = m_q - m_0, \quad O = m_o - m_0, \qquad A = D + Q, \quad S = D + O.$$

Inference weights the ten subject lexemes equally after averaging their sentence combinations. The fourteen contrasts comprise eight simple operation/dose differences, two interactions, and four low-dose PP-cue contrasts. Exact two-sided tests enumerate all $2^{10} = 1,024$ subject-sign assignments and apply Holm-14. Individual 95% intervals use 20,000 subject-cluster bootstrap draws, providing marginal coverage for each contrast. All 14 adjusted probabilities are $.027344$, partly reflecting the exact test's finite resolution (Table S14).

The Hebrew-derived advantage increases when both edits move from lower to higher strength, for attenuation and selectivity alike (Table S14). The cross-language amplitude difference therefore changes with intervention strength. One subject lexeme reverses the attenuation difference at the lower strength, but omitting any subject or target preserves the average contrast. These deletion comparisons assess the same contrast on overlapping materials.

Across sentence configurations, English-derived effects retain more of the Hebrew-derived response at the common lower strength: ratios are .867 for attenuation and .819 for selectivity, compared with .773 and .746 at the higher strength. The English edit used in the original transfer is unchanged; lowering the Hebrew reference changes what it is compared against. Thus the earlier near-30% normalized ratio and these common-strength raw ratios answer different questions. Two strengths establish that the cross-language amplitude comparison is conditional on intervention strength.

Table S14: Fourteen primary Hebrew comparisons, giving ten subject lexemes equal weight. Native and Source denote Hebrew- and English-derived edits; L/H denote lower/higher strengths. Intervals are individual subject-bootstrap intervals. Exact two-sided sign flips use all $2^{10} = 1,024$ assignments; Holm correction covers all 14 contrasts. Each interval gives marginal coverage for its contrast.

Effect	Comparison	Estimate	95% interval	p	p_H
A	Native-source, L	0.020728	[0.0100, 0.0326]	0.005859	0.027344
A	Native-source, H	0.096794	[0.0688, 0.1249]	0.001953	0.027344
A	Source, H-L	0.196755	[0.1474, 0.2508]	0.001953	0.027344
A	Native, H-L	0.272821	[0.2085, 0.3426]	0.001953	0.027344
A	(Native-source), H-L	0.076066	[0.0550, 0.0961]	0.001953	0.027344
A	PP Source, mismatch-match	0.059904	[0.0370, 0.0851]	0.001953	0.027344
A	PP Native, mismatch-match	0.072857	[0.0458, 0.1031]	0.001953	0.027344
S	Native-source, L	0.028073	[0.0171, 0.0399]	0.001953	0.027344
S	Native-source, H	0.108094	[0.0760, 0.1392]	0.001953	0.027344
S	Source, H-L	0.192439	[0.1430, 0.2475]	0.001953	0.027344
S	Native, H-L	0.272460	[0.2056, 0.3447]	0.001953	0.027344
S	(Native-source), H-L	0.080021	[0.0571, 0.1018]	0.001953	0.027344
S	PP Source, mismatch-match	0.064045	[0.0390, 0.0917]	0.001953	0.027344
S	PP Native, mismatch-match	0.071724	[0.0450, 0.1009]	0.001953	0.027344

C.11 Shared sentence sensitivities and the effects of comparison edits

The two language-derived edits agree strongly on sentence susceptibility: raw rank correlations exceed .91 for both endpoints at either strength. The subject-edit contributions share this ordering ($\rho = .92$ and .96 at the two strengths), as do the effects of editing another position (.93 at both); alternative-direction effects correspond at $\rho = -.05$ and $-.01$. Thus the selective number effect combines shared sentence susceptibility with greater disruption at the subject than under the direction and position comparisons (Table S15). Ranking and selectivity describe distinct properties of the transferred influence.

Normalization can separate response ordering from linear similarity in effect size. For simple agreement at L, normalized amplitudes have a linear correspondence of $r = .028$, while sentence rankings remain related ($\rho = .748$). A near-zero normalization reference magnifies one negative English-derived response and strongly affects the linear correlation. The observation remains in every relevant summary, with its components available online. The resulting contrast explains why similar rankings can coexist with dissimilar normalized amplitudes.

Both operations include negative sentence responses at both strengths. Transfer therefore preserves a broad ordering of susceptibility while allowing individual responses to reverse. Online data retain every response, construction, and normalization choice. The all-material and prepositional-phrase summaries in the cue comparison both use the same 60 configurations.

C.12 Intervening noun number changes the response to intervention

The intervening-noun comparison changes number context while preserving the position of the edit and target. Ten subject lexemes and seven intervening nouns, each in singular and plural, and seven verbs form 60 configurations selected before outcomes. Crossing subject and intervening-noun number yields 240 items, four per configuration, and 120 intervening-noun contrasts between items that differ only in that noun's number; each contrast preserves target token identities and indices, sequence length, subject and off-position indices, and intervening distance. They are new combinations of vocabulary from the existing pool.

For example, "the waiter.SG near the composer.SG travel.SG" is compared with "the waiter.SG near the composer.PL travel.SG" and the two corresponding plural-subject forms. These glosses illustrate the manipulation; actual Hebrew strings retain the benchmark's separate article tokens and spacing. Singular/plural head-final positions can differ under Qwen-7B tokenization, but each

Table S15: Raw subject, orthogonal-comparison and off-position components for Hebrew-derived (Native) and English-derived (Source) edits. D is baseline minus subject edit; Q/O are comparison minus baseline. Thus A=D+Q and S=D+O. Hebrew A uses the perpendicular-direction reference; English A in Table S10 uses the unedited response. The difference in dose effects is the native-minus-source difference in high-minus-low responses. Subject-edit effects contribute to both contrasts. Cue results at the higher strength lie outside the prespecified fourteen-comparison family.

Comparison	Dose	Direction	D	Q	O	A	S
Native minus source	L		0.032064	-0.011336	-0.003991	0.020728	0.028073
Native minus source	H		0.114629	-0.017835	-0.006534	0.096794	0.108094
High minus low		Source	0.198512	-0.001757	-0.006073	0.196755	0.192439
High minus low		Native	0.281077	-0.008255	-0.008617	0.272821	0.272460
Difference in dose effects			0.082565	-0.006499	-0.002544	0.076066	0.080021
Mismatch minus match	L	Source	0.068427	-0.008523	-0.004382	0.059904	0.064045
Mismatch minus match	L	Native	0.076118	-0.003261	-0.004394	0.072857	0.071724
Mismatch minus match	H	Source	0.133187	-0.014233	-0.007113	0.118954	0.126074
Mismatch minus match	H	Native	0.146475	-0.005325	-0.007131	0.141151	0.139344

Table S16: Reciprocal subject-number and intervening-noun-number comparisons. Entries give raw margins, weighting subject lexemes equally; negative baselines favor the disagreeing target. Native/Source identify Hebrew-/English-derived edits; L/H identify lower/higher strengths. The prespecified fourteen-comparison family contains the lower-strength cue contrasts; high-dose and subject-number-specific contrasts are reported outside it.

Subject	Cue	Direction	Dose	Baseline	A	S	D
PL	Match	Source	H	1.313600	0.120850	0.102364	0.117835
PL	Match	Source	L	1.313600	0.048993	0.036726	0.046161
PL	Match	Native	H	1.313600	0.159112	0.153174	0.175745
PL	Match	Native	L	1.313600	0.060996	0.057695	0.071476
PL	Mismatch	Source	H	-1.479399	0.182526	0.157218	0.180818
PL	Mismatch	Source	L	-1.479399	0.074816	0.058436	0.072872
PL	Mismatch	Native	H	-1.479399	0.213590	0.216200	0.245640
PL	Mismatch	Native	L	-1.479399	0.081059	0.083110	0.101115
SG	Match	Source	H	2.729035	0.245775	0.225450	0.246890
SG	Match	Source	L	2.729035	0.113068	0.100685	0.113766
SG	Match	Native	H	2.729035	0.310187	0.284071	0.310800
SG	Match	Native	L	2.729035	0.131345	0.115343	0.131647
SG	Mismatch	Source	H	0.168567	0.422008	0.422743	0.450281
SG	Mismatch	Source	L	0.168567	0.207053	0.207064	0.223907
SG	Mismatch	Native	H	0.168567	0.538011	0.499735	0.533856
SG	Mismatch	Native	L	0.168567	0.256996	0.233375	0.254243

intervening-noun contrast preserves the relevant positions. Noun number and token identity change together. There is no Hebrew bridge edit.

Conflicting intervening-noun number amplifies both language-derived edits at the fixed lower strength. Attenuation and selectivity increase in all four contrasts of the prespecified fourteen-comparison family. This increase is already present relative to the unedited response, while alternative-direction and alternative-position changes partly offset it. Number context therefore changes susceptibility to subject editing itself, beyond its effect on the comparison edits. The prespecified fourteen-comparison family contains the four lower-strength cue contrasts; higher-strength increments and the cue-by-operation, cue-by-strength, and three-way interactions are reported outside that family.

Original agreement difficulty and sensitivity to intervention respond differently to subject number (Table S16). Conflicting nouns reduce unedited agreement more for plural subjects, with some

contexts already favoring the disagreeing verb, yet amplify the edit more for singular subjects. Contextual susceptibility therefore carries information beyond the model's initial preference for the correct form. The mismatch-minus-match difference combines possible protection by matching number and interference by conflicting number, because the design varies neither distance nor a neutral noun-number condition.

C.13 Design chronology and comparison scope

The held-out English and common-strength Hebrew designs were fixed after the earlier findings and before their own outcome analyses, and the confined-edit comparison (Appendix C.5) and the shared/model-specific comparison (Appendix D.6) were fixed in the same way, together with how their outcomes would be read, before their outcomes were computed. The designs remained fixed during scoring and fold-level inference. Their prespecified comparisons are distinguished from the original English-to-Hebrew transfer.

The present and earlier English evaluations used single-precision arithmetic (float32). The original Hebrew analysis specified half precision (float16) on MPS, with realized precision unrecorded. Operator fitting, weighting and outcome definitions also differ across these evaluations, so comparisons reflect these methodological choices jointly. The anonymous repository provides exact materials, model settings, and condition-level scores for each comparison.

D Brain–model correspondence and neural sampling

D.1 SEEG acquisition and listening procedure

SEEG recordings were obtained during clinical evaluation for epilepsy while participants listened to Chinese sentence streams. Participants received oral instructions to listen attentively. Each sentence comprised four syllables, each lasting 250 ms; the sentence-listening condition included 40 trials. To reduce trial-onset responses, analysis omitted the initial four-character item and retained eight sentences (32 syllables) per trial, yielding 320 sentence occurrences per participant.

The signals used for analysis were sampled at 512 Hz for 20 participants and 2048 Hz for one participant; feature extraction used each recording's sampling grid. The study received institutional review board approval. Participants gave informed consent after receiving information about auditory-stimulation risks, and compensation was prorated to task completion.

D.2 Participants, signals, and anatomical coverage

Twenty-one participants meet the prespecified timing, signal and channel criteria, including the exact mapping of 40 trials with eight analyzed four-character sentences each. Each sentence contains two two-character words. Six of the 27 available participant records lack the required stimulus mapping or trial count.

The primary analysis includes all 1,407 channels with finite measurements and anatomical labels from the 21 participants. Coordinates locate 1,341 channels from 20 participants; the remaining 66 channels have anatomical labels but no available coordinates. Statistical inference therefore includes all 21 participants, while the surface image shows recording coverage for 20. The image maps recording coverage through channel coordinates; statistical effects are summarized separately across participants.

Two signal summaries define the neural targets. The low-frequency summary uses a fourth-order zero-phase 4-Hz low-pass filter, a -400 to -100 ms baseline, and mean absolute amplitude in four consecutive 250-ms windows beginning 50 ms after each sentence onset. The high-gamma summary averages baseline-relative Morlet power in dB over 70–150 Hz, using ten 50-ms bins from 50 to 550 ms after each of the item's two word onsets. Low-frequency amplitude and high-gamma power summarize complementary temporal aspects of the measured response; each can reflect multiple linguistic processes. An integrated target combines separately centred and Frobenius-normalized branch kernels with equal weights.

Table S17: Family-weighted additional same-sentence correspondence (observed correlation minus the trial-reassignment mean), with unadjusted whole-trial randomization probabilities. Integrated/Subspace is the primary comparison. BH correction is applied within the family of 81 model-level comparisons of neural targets and representations.

Neural target	Concept excess (p)	Subspace excess (p)	Rhythm excess (p)
Integrated	.001495 (.100)	.001894 (.072)	.001418 (.100)
High-gamma	.001645 (.088)	.002710 (.028)	.001162 (.167)
Low-frequency	.000737 (.255)	.000284 (.413)	.000644 (.262)

D.3 Held-out correspondence and its randomization reference

Held-out trials test whether model and neural representations preserve similar distinctions among sentence occurrences. For each participant, two exchanged 20-trial halves define training and evaluation; neural columns are standardized using training means and scales. Four complete-trial folds within each training half select model PCA rank from $\{1, 2, 4\}$ and ridge-whitening coefficient from $\{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 100, 1000\}$. The selected transformation is evaluated on the untouched half. Correspondence is Pearson correlation of off-diagonal upper-triangle entries in the centred model and neural kernels, distinct from the atlas's Spearman statistic.

Whole-trial reassignment separates correspondence to the same stimuli from resemblance that survives their mismatching. The primary comparison relates integrated neural geometry to model Subspace, weighting the five model families and participants equally. Each of 1,000 null draws permutes complete eight-sentence sequences within the outer halves and repeats inner rank/ridge selection; PCA bases remain fitted to model training data. The one-sided plus-one probability is $(1 + b)/1001$. Participant intervals use 10,000 resamples for the fixed stimuli. Model max-T, the FDR family of 81 model/target/representation combinations, and two aggregate Holm-adjusted secondary contrasts retain separate inference scopes.

Secondary comparisons locate the additional stimulus correspondence among representations and signal summaries. One asks whether Subspace correspondence on the integrated target exceeds the mean of Concept and Rhythm; the other asks whether Concept-versus-Rhythm correspondence differs between the two neural targets. These comparisons ask which representational view captures distinctions among the same stimuli in models and brains; the number experiments examine grammatical influence on model predictions.

D.4 Additional same-sentence correspondence and its reference scale

Observed brain–model correspondence exceeds the mean obtained after whole-trial reassignment. The observed correlation is $.018529$, compared with $.016634$ after whole trials are mismatched while sequence positions are preserved, yielding additional same-sentence correspondence of $.001894$. The increment is positive in fourteen participants and negative in seven. Its participant-bootstrap interval $[.000092, .003825]$ is above zero for the fixed stimuli, and the whole-trial randomization gives $p = .071928$. These comparisons separate two scientific questions: whether the sampled participants have a positive average increment, and how the original sentence correspondence compares with alternative trial assignments.

All nine models have positive estimates of additional same-sentence correspondence on the common neural sample (Table S18), although their magnitudes differ across the five related families. All nine estimates describe the same participants, yielding a cross-model comparison on one neural sample. Among secondary comparisons, Subspace with high-gamma responses has the largest unadjusted association; BH correction covers the family of 81 model-level comparisons (Table S17).

Neural reliability places the additional correspondence on the scale of distinctions reproducible across participants. Against the integrated lower correlation reference ($.263660$; upper reference $.378976$), the increment is about $.72\%$, with a 90% interval of $.14\%$–1.32%. This lies within the prespecified 10% region and the 5% and 20% sensitivity regions. This percentage is a ratio of two correlations: the same-sentence increment divided by the lower interparticipant reliability reference.

Table S18: Per-model correspondence and neural reliability references. Excess is the observed correlation minus the trial-reassignment mean. The 95% intervals resample participants with stimuli fixed. Probabilities use whole-trial randomization; max-T adjusts across nine models sharing the same neural sample. Lower and upper reliability bounds give interparticipant reference levels.

Model	Observed	Null	Excess	95% interval	p	p_{maxT}
Llama-8B	0.019179	0.016756	0.002423	[-0.00105, 0.00597]	0.100899	0.370629
Llama-ZH	0.016110	0.014122	0.001988	[-0.00144, 0.00572]	0.120879	0.471528
DS-Llama	0.011597	0.007959	0.003638	[0.00066, 0.00720]	0.010989	0.050949
Qwen-7B	0.015348	0.012835	0.002513	[-0.00049, 0.00539]	0.079920	0.325674
Qwen-7B-It	0.013278	0.011418	0.001859	[-0.00093, 0.00454]	0.132867	0.481518
Qwen-14B	0.017478	0.015101	0.002377	[-0.00120, 0.00612]	0.089910	0.355644
Mistral	0.012744	0.011438	0.001306	[-0.00116, 0.00364]	0.221778	0.657343
Gemma-9B	0.039261	0.039175	0.000086	[-0.00684, 0.00880]	0.484515	0.904096
GPT-2-ZH	0.009642	0.006495	0.003147	[-0.00014, 0.00652]	0.022977	0.099900

Neural reliability reference	Lower bound	Upper bound
High-gamma	0.203453	0.337555
Low-frequency	0.139237	0.295721
Integrated	0.263660	0.378976

D.5 Participant, stimulus, and analysis sensitivity

The positive average persists when any one participant, model family, or trial is omitted, with excess ranges .001348–.002304 across participants, .001581–.002346 across model families and .001035–.003037 across trials (Table S19).

Adjusting for sequence structure, lexical properties, and semantic grouping leaves positive point estimates, with several participant intervals including zero. Sequence terms cover trial identity, position identity, and position difference. Lexical terms cover identity, constituent log-frequency differences and missingness; semantic-cluster identity accounts for shared meaning groups, and an additional variant includes token-count terms. The increment remains positive across the examined ranks and adjustments. These supplementary randomizations hold the observed rank and regularization fixed; the primary test repeats selection for each permutation.

Training-based channel selection has different effects across signal summaries on the smaller matched samples (Table S20). In the 18-participant low-frequency comparison, selection increases the same-sentence increment: the unadjusted one-sided trial comparison gives $p = .026$, while the participant interval for the difference includes zero. In the eight-participant high-gamma comparison, selected channels have a smaller increment, again with a participant interval spanning zero. The selected-minus-all participant intervals include zero for all three targets. Channel selection therefore shows signal-dependent point estimates with uncertainty across participants; the full eligible set remains the primary analysis, and the anatomical figure shows its sampling coverage.

D.6 Shared and model-specific parts of the representation

The comparison asks whether the part of a model's sentence representation predicted from other model families has greater additional same-sentence correspondence than the model-specific remainder. Both parts enter the same held-out analysis. For each model, the Subspace representation of the 320 sentence occurrences (Appendix D.3) is split into two parts of equal dimension. The shared part is the linear prediction of the model's representation from the representations of the models in the other four families: for each of those models the first eight principal components are taken, averaged within each family, concatenated across families and reduced again to eight components, and a ridge regression with unit penalty and an unpenalized intercept maps these components to the model's representation. The prediction is the shared part and the remainder the model-specific part; each is then represented by its first four principal components, so both enter the correspondence analysis with four dimensions, the largest rank among those the analysis selects from. Reducing the pooled

Table S19: Neural correspondence when sampled units are omitted or analysis choices vary. Omission ranges describe sensitivity to one participant, model family, or trial; the lower block uses participant-bootstrap intervals for fixed stimuli. The primary analysis reselects rank and regularization under each reassignment; supplementary tests keep the fitted transformation fixed.

Deletion analysis	Excess range	Summary
Omit one participant	0.001348–0.002304	Omission range
Omit one model family	0.001581–0.002346	Omission range
Omit one trial	0.001035–0.003037	Omission range
Analysis	**Excess**	**95% participant interval**
Primary analysis	0.001894	[0.000092, 0.003825]
Unadjusted, fixed transformation	0.002131	[0.000242, 0.004168]
Linguistic covariates	0.002116	[-0.000039, 0.004325]
+ token count	0.002092	[0.000049, 0.004107]
Fixed rank 1	0.000806	[-0.000895, 0.002532]
Fixed rank 2	0.001904	[-0.000199, 0.004107]
Fixed rank 4	0.001743	[-0.000236, 0.003809]

Table S20: All-channel and training-selected channel comparisons on endpoint-specific matched samples. Probabilities are unadjusted one-sided whole-trial randomization tests; 95% intervals resample participants for the fixed stimuli. Selected-minus-all intervals include zero for all three targets.

Target	Comparison	N	Excess	95% interval	p
High-gamma	all channels	8	0.006449	[0.00246, 0.01030]	0.006993
High-gamma	train-only selected	8	0.003371	[-0.00069, 0.00772]	0.079920
High-gamma	selected - all	8	-0.003078	[-0.00769, 0.00150]	0.943057
Integrated	all channels	8	0.002629	[-0.00044, 0.00543]	0.127872
Integrated	train-only selected	8	0.002006	[-0.00273, 0.00642]	0.175824
Integrated	selected - all	8	-0.000623	[-0.00386, 0.00218]	0.641359
Low-frequency	all channels	18	-0.000762	[-0.00289, 0.00164]	0.704296
Low-frequency	train-only selected	18	0.001565	[-0.00144, 0.00444]	0.118881
Low-frequency	selected - all	18	0.002328	[-0.00029, 0.00493]	0.025974

components to eight, twice the four retained per part, keeps the shared parts of models from the same family distinct; with equal ranks the shared part is the same object for every model of a family, because the correspondence measure is unchanged by invertible linear maps of a representation. Standardization, principal components and the regression are fitted on the 160 training items of each half-trial split and applied to all 320 items, and each half is scored only by the split fitted on its own training items, so held-out items never inform the split. The correspondence analysis, the 1,000 whole-trial reassignment draws, the same draws for both parts, the 10,000-draw participant bootstrap, the use of all eligible channels and the family weighting are those of Appendix D.3; before the split was introduced, the implementation reproduced the published values of Appendix D.4 exactly (observed .018529, reassignment mean .016634, increment .001894, $p = .071928$). Concept is analyzed in the same way as a secondary representation, and the low-frequency and high-gamma targets as secondary targets; Rhythm has four features and is not split. The comparison and its reading were fixed before any result was computed: the primary quantity is the family-weighted difference between the shared and the model-specific increment on the integrated target, the sign of the difference is left open when its 95% participant interval includes zero, and a difference larger than one tenth of the lower reliability reference is excluded when the 90% interval lies inside that region, as in Appendix D.4.

Mean additional same-sentence correspondence is .00236 for shared and .00060 for model-specific parts, weighting families equally. Their difference is .00176 (95% participant interval $[-.00163, .00526]$; whole-trial reassignment $p = .137$; Table S21). The 90% interval,

Table S21: Shared and model-specific parts of the Subspace representation and their additional same-sentence correspondence on the integrated target. Each value is the observed correlation minus the whole-trial reassignment mean; Difference is shared minus specific. Intervals resample the 21 participants with stimuli fixed (10,000 draws); p is the one-sided whole-trial reassignment probability of the difference (1,000 draws), and q its BH-corrected value within the 54 model-by-target-by-representation comparisons. The all-model row weights the five families equally; its 90% interval, $[-.00112, .00471]$, lies inside one tenth of the lower reliability reference of .264. The lower block gives the family-weighted difference for each representation and neural target; Holm covers the three secondary comparisons; the two remaining Concept rows carry no correction.

Model	Shared	Specific	Difference	95% interval	p	q
Llama-8B	.00205	.00397	−.00192	[−.00669, .00261]	.783	.900
Llama-ZH	.00308	.00096	.00211	[−.00228, .00617]	.180	.545
DS-Llama	.00179	.00173	.00006	[−.00494, .00538]	.487	.764
Qwen-7B	.00234	−.00089	.00323	[−.00291, .00905]	.096	.398
Qwen-7B-It	.00170	−.00127	.00298	[−.00276, .00885]	.089	.398
Qwen-14B	.00289	.00051	.00238	[−.00398, .00856]	.168	.545
Mistral	.00246	.00106	.00141	[−.00342, .00628]	.259	.607
Gemma-9B	.00159	.00095	.00064	[−.00776, .00769]	.442	.755
GPT-2-ZH	.00312	−.00067	.00379	[−.00028, .00792]	.046	.276
All models	.00236	.00060	.00176	[−.00163, .00526]	.137	—

Representation	Neural target	Shared	Specific	Difference	95% interval	p	p_H
Subspace	Integrated	.00236	.00060	.00176	[−.00163, .00526]	.137	—
Subspace	High-gamma	.00223	.00017	.00206	[−.00188, .00584]	.129	.387
Subspace	Low-frequency	.00155	.00041	.00115	[−.00199, .00447]	.249	.498
Concept	Integrated	.00124	.00098	.00026	[−.00296, .00347]	.422	.498
Concept	High-gamma	.00195	.00046	.00149	[−.00171, .00476]	.148	—
Concept	Low-frequency	.00014	.00082	−.00069	[−.00325, .00192]	.698	—

$[-.00112, .00471]$, lies inside one tenth of the lower reliability reference (.264; the region is $\pm.0264$) and inside one twentieth of it. The sign is left open, and the two parts show comparable same-sentence correspondence within this margin. Thirteen of the 21 participants have a positive difference; omitting any one participant leaves it between .00080 and .00254, and omitting any one model family between .00125 and .00217. Eight of the nine models share the direction of the average, Llama-8B the exception; the smallest Holm-corrected per-model probability is .41, and the smallest BH-corrected value in the 54 model-by-target-by-representation comparisons is $q = .216$. The difference is .00206 on the high-gamma target and .00115 on the low-frequency target, and .00026 for Concept on the integrated target. Predicting the shared part from all other models, the same family included, gives .00178; ridge penalties of 0.1 and 10 give .00147 and .00097; pooling four or sixteen components in place of eight gives −.00060 and −.00011, the coincidence of shared parts within a family that the eight-component choice avoids. Adjusting for trial and position, lexical identity and frequency, semantic cluster and token counts leaves the difference at .00149 to .00165. A version with two components per part is not estimable, because the correspondence analysis requires at least four dimensions per representation. Excluding Llama-8B, whose Llama 3 generation is identified by configuration alone (Table S3), leaves the difference at .00196 when the model is dropped from the average and .00137 when the shared parts are refitted without it.