Attention-Side Transformer Computations Organize Shared Multilingual Brain Alignment during Naturalistic Story Listening

Ali Uyar

Independent Researcher

Abstract

Most brain-language-model alignment studies rely on returned hidden states, leaving unclear which internal transformer computations carry cross-lingually shared signal during naturalistic comprehension. We analyzed English, French, and Chinese listeners from the Le Petit Prince multilingual naturalistic fMRI corpus (94 participants; 9 runs per language cohort) using ROI-average encoding models built from intermediate states of XLM-R-base and the encoder of NLLB-200-distilled-600M. For each sentence span and target language, we decomposed state representations into a leave-target-out SHARED component and an orthogonalized SPECIFIC residual, then contrasted attention-side and FFN-side contributions within semantic and auditory ROI families. The planned semantic FFN-preference hypothesis failed uniformly: all six semantic model-language contrasts were negative (mean effect −0.0083; 0/6 Holm-significant in the planned direction), and supplementary reverse-direction tests favored attention-side states in all six cells. By contrast, the auditory attention-preference hypothesis was positive in 5/6 cells and Holm-significant in 4/6 (mean effect 0.0033). Representative winners concentrated in ATTN and POST-ATTN rather than FFN states, semantic representative settings showed a larger mean content bias than auditory settings (0.663 vs. 0.610), and deleting top-attributed words hurt held-out prediction more than matched-random deletion in all 20 summary rows (mean top-k minus random difference 0.0381). These results do not support a late FFN-centered account of shared multilingual brain alignment. Instead, they indicate that the shared multilingual advantage is organized primarily around attention-side and post-attention computations, with the clearest confirmatory support in auditory cortex and a consistent semantic shift away from FFN-side states.

Keywords: Naturalistic fMRI, multilingual language models, transformer internals, brain encoding, token attribution, computational neuroanatomy

1. Introduction

Naturalistic neuroimaging has made it possible to study language comprehension under continuous, ecologically valid stimulation rather than only under isolated trial structures [1–3]. In parallel, modern transformer language models have become useful computational probes for human language processing, because their internal representations often predict neural responses in text and speech paradigms [4–7]. Most of this literature, however, still treats a model as a sequence of returned hidden states. That abstraction is useful, but it blurs a mechanistically important question: which internal computations within the transformer block carry the signal that aligns with the brain?

That question is particularly sharp for multilingual models. Multilingual encoders such as XLM-R and NLLB are explicitly trained to support cross-lingual transfer [8, 9]. They therefore provide a natural test bed for separating structure that is shared across languages from structure that remains language specific. In the present work, we operationalize that contrast using a leave-target-out SHARED representation and an orthogonalized target-language SPECIFIC residual. The resulting factorization asks not whether a model predicts the brain in general, but whether the part of the representation that generalizes across languages is carried by the same internal computations as the part that remains language specific.

The original sequel plan motivating this project predicted that semantic cortex would preferentially express the shared multilingual advantage in later FFN-side computations, whereas auditory cortex would show a stronger attention-side pattern. The present analyses do not support that exact asymmetry. Instead, the data support a more interesting and more defensible result: the multilingual shared advantage is computationally structured, but the strongest structure is attention-linked rather than late-FFN dominated. In the semantic ROI family, every planned FFN-versus-attention contrast goes in the opposite direction. In the auditory ROI family, attention-side preference is replicated in most model-language cells and survives familywise correction in four of six.

This paper therefore takes the completed internal-state evidence as primary and treats the work as a standalone mechanistic study. The central question is not whether a previous bridge analysis can be refreshed exhaustively, but whether the finished mechanistic evidence supports a publishable account of how multilingual transformer computations map onto naturalistic language responses in the human brain. The answer is yes. Using ROI-average encoding analyses in the Le Petit Prince multilingual naturalistic fMRI corpus, we show that the shared multilingual advantage is not a monolithic hidden-state

property. Instead, it is concentrated in attention-side and post-attention computations, with the clearest confirmatory support in auditory cortex, a robust semantic reversal away from FFN-side states, and convergent token-level evidence from attribution and deletion validation.

2. Results

2.1. The shared multilingual advantage is structured across internal transformer states

The state sweep reveals a pronounced state-by-depth organization of the SHARED minus SPECIFIC advantage across both model families and all three language cohorts (Fig. 1). This pattern is not flat and it is not confined to a single late hidden state. Instead, the heatmaps show consistent internal structure across block depth, with warm bands concentrated in ATTN, POST-ATTN, and, in some settings, OUTPUT states. Across the full representative-state summary, semantic winners comprised four ATTN states, one POST-ATTN state, and one OUTPUT state, with no semantic FFN winners at all; auditory winners comprised one ATTN state, three POST-ATTN states, and two OUTPUT states. The internal-state pattern is therefore already inconsistent with a strong late-FFN account before formal hypothesis testing.

This computational structure is visible in both ROI families but differs in strength and localization. The auditory family exhibits the clearest concentration of positive SHARED minus SPECIFIC effects in ATTN and POST-ATTN states, especially in XLM-R. The semantic family shows a broader internal-state pattern, but crucially that pattern is shifted away from the planned FFN-side explanation rather than toward it. In other words, the semantic result is not a null. It is an organized reversal.

2.2. The planned semantic FFN hypothesis reverses, whereas the auditory attention hypothesis is supported

The primary confirmatory tests are summarized in Fig. 2. The planned semantic hypothesis (H1: FFN-side > attention-side within semantic ROIs) failed in the strongest possible directional way: all six model-language contrasts were negative, all six confidence intervals lay below zero, and none was significant in the planned positive direction. The mean H1 effect across cells was -0.0083. Thus, the semantic-family data do not merely withhold support from FFN-side dominance; they systematically contradict it.

Because the uniform H1 reversal is itself scientifically informative, we quantified the opposite direction in an explicit post hoc analysis of the same subject-level effects (Supplementary Table B.4; Supplementary Figure A.6). In that exploratory reverse-direction test, all six semantic cells favored attention-side over FFN-side states after Holm correction. We emphasize that these reverse-direction tests are post hoc rather than preregistered, but they show that the semantic reversal is not a weak descriptive fluctuation.

The auditory hypothesis (H2: attention-side > FFN-side within auditory ROIs) was much closer to the originally planned story. Five of six model-language cells were positive, and four of six survived Holm correction. The clearest support came from XLM-R in all three language cohorts and from NLLB in Chinese. NLLB English was negative and NLLB French was near zero, so the auditory pattern is not universal; however, it is strong enough to support a positive family-level claim rather than a merely descriptive one.

Taken together, the primary tests support a sharper and more realistic conclusion than the planned one. The multilingual shared advantage is not concentrated in late semantic FFN computations. Instead, it is attention-linked, with strong confirmatory support in the auditory family and a robust semantic reversal away from FFN-side dominance.

2.3. Representative ROI profiles show interpretable regional heterogeneity

Representative ROI curves make the family-level results anatomically interpretable (Fig. 3). In angular gyrus and pars triangularis, SHARED minus SPECIFIC effects remain positive across much of the depth axis, but the highest curves are predominantly ATTN POST-ATTN, or OUTPUT rather than FFN. In posterior superior temporal gyrus and Heschl's gyrus, the curves are also consistently positive, but the auditory pattern is more strongly attention-linked and shows larger early-to-mid depth separation.

This matters for interpretation. A simple semantic-versus-auditory dissociation would imply that the semantic family should look late and FFN-centric while the auditory family looks early and attention-centric. The actual data are more nuanced. Both families show structured positive shared signal. The strongest family-level distinction is that the auditory pattern is confirmed by the pre-specified attention-side hypothesis, whereas the semantic pattern organizes around an unplanned but internally consistent reversal toward earlier computations.

2.4. Token-level evidence is consistent with a content-sensitive semantic signal

Token-level analyses were restricted to representative winner settings and should therefore be interpreted as secondary mechanistic support rather than primary evidence. With that caveat, the token results were coherent. Semantic representative settings showed a larger mean content bias than auditory representative settings (0.663 vs. 0.610), indicating that the semantic-family shared signal was more concentrated on content-bearing words than the auditory-family shared signal. The token-class mass summary is shown in Fig. 4; an illustrative pair of multilingual examples is provided in Supplementary Figure A.7.

The token examples are informative for narrative purposes but are not themselves inferential. The stronger evidential role is played by the aggregate token-class summary and, especially, by deletion validation.

2.5. Deletion validation confirms that top-attributed words are functionally relevant

Deletion validation provides the most direct secondary test of whether the attribution signal is functionally meaningful. Across 10 representative settings and deletion depths $k \in \{1, 2\}$, deleting top-attributed words reduced held-out predictive performance more than matched-random deletion in all 20 summary rows

(Fig. 5). The mean top-k minus random difference was 0.0381, with slightly larger average effects in semantic than auditory settings (0.0402 vs. 0.0350). Deeper deletion also strengthened the effect: the mean difference rose from 0.0294 at $k = 1$ to 0.0469 at $k = 2$.

These deletion results do not establish that the token-level attribution exhausts the underlying computational mechanism. They do, however, show that the ranked words identified by the attribution procedure are not purely cosmetic. The words most implicated by the model-side attribution are also the ones whose removal most strongly disrupts predictive performance.

3. Discussion

The strongest defensible conclusion from the present analyses is that shared multilingual brain alignment is attention-linked rather than late-FFN dominated. This conclusion is stronger than the original planned semantic headline and more interesting scientifically. The planned semantic FFN hypothesis failed in all six model-language cells, the family's representative states were almost entirely attention-side or post-attention, and reverse-direction post hoc tests favored attention-side states throughout. Meanwhile, the auditory attention hypothesis was positively supported in most cells and survived correction in four of six.

This pattern matters for how multilingual transformer-brain alignment is interpreted. A common simplification is to treat a model's returned hidden states as the relevant computational object. Our results argue that this is too coarse. The explanatory signal is distributed across internal transformer computations, and the part of the representation that is shared across languages is not captured equally by all of them. In the present data, the shared signal is most clearly expressed in attention-linked computations. That does not imply that FFN states are irrelevant; it implies that the critical cross-linguistically shared variance is already prominent before or around the attention-residual transition and is not best summarized as a late FFN-only code.

The auditory family provides the clearest confirmatory support for this interpretation. This is notable because auditory and superior temporal regions are sometimes treated as a nuisance complication whenever shared cross-linguistic signal appears outside classical semantic ROIs. Here, the auditory result is not merely residual shared variance. It is the strongest pre-specified effect in the paper. The semantic family is more subtle: its signal is robustly positive in terms of SHARED minus SPECIFIC prediction, but the internal-state structure points away from late FFN dominance. That shift makes the paper more mechanistic, not less. It suggests that the semantic-family shared signal may depend on computations that are already substantially resolved at the attention and post-attention stages.

The token-level results are consistent with this interpretation. Semantic representative settings were more content-biased than auditory settings, and top-attributed deletions consistently outperformed matched-random deletions. We treat these findings as secondary, because representative-state selection preceded the token analyses and because the family-level content-bias contrast was not one of the primary confirmatory tests. Even so, the combination of descriptive token-class structure and uniformly positive deletion validation makes the token section substantively useful rather than decorative.

More broadly, the present paper illustrates a productive way to use multilingual language models in cognitive neuroscience. The goal need not be to maximize raw encoding performance or to claim that one model is "the" brain model. Instead, multilingual encoders can be used as computational probes that factorize shared versus language-specific information and localize those factors to internal computations. On that criterion, the present analyses support a clear message: the multilingual shared advantage is computationally organized, and the organization is predominantly attention-linked.

4. Limitations and scope

Several limitations should temper interpretation. First, the study uses ROI-average naturalistic fMRI encoding rather than voxelwise analyses, so the reported effects characterize region families and representative ROIs rather than fine-grained cortical topographies. Second, the strongest semantic attention-side claim is post hoc: the pre-specified semantic FFN hypothesis failed uniformly, and the reverse-direction semantic tests were conducted only after inspecting that result. We regard those reverse-direction tests as legitimate and informative, but they remain exploratory. Third, the token-level analyses are based on representative winner settings rather than a full all-state token sweep, and the content-bias contrast is descriptive rather than confirmatory. Fourth, the paper is intentionally written as a standalone mechanistic study and therefore does not rely on unfinished output-state bridge-refresh artifacts. That choice strengthens the internal-state claim but leaves continuity-style bridge arguments outside the present manuscript's scope. Finally, the study covers two multilingual encoder families and one naturalistic corpus; broader generalization across architectures, modalities, and datasets remains open.

5. Methods

5.1. Dataset and sample

All analyses used the Le Petit Prince multilingual naturalistic fMRI corpus (LPPC-fMRI) [3], a public dataset in which participants listened to the same audiobook in their native language. The analyzed sample contained 94 participants distributed across three language cohorts: 31 English listeners, 28 French listeners, and 35 Chinese listeners. Each cohort contributed nine canonical runs. The repetition time was 2.0 s in all three cohorts.

5.2. Sentence spans, aligned triplets, and model states

The linguistic unit of analysis was the aligned sentence span. For every aligned multilingual triplet, we extracted sentence-pooled transformer representations from XLM-R-base and from the encoder of NLLB-200-distilled-600M [8, 9]. Four internal states were retained densely for each of the 12 encoder blocks: ATTN, POST-ATTN, FFN, and OUTPUT. Bundle-level model inventory files reported that output-state equivalence checks passed for both models.

5.3. SHARED and SPECIFIC factorization

For each sentence span s, target language ℓ, and internal state, let $\tilde{h}_{s,\ell}$ denote the pooled and normalized target-language vector. The leave-target-out shared component was defined as the average of the non-target language vectors,

$$u_{s,\ell} = \frac{1}{L-1} \sum_{j \neq \ell} \tilde{h}_{s,j}, \tag{1}$$

where $L = 3$ languages. The raw target-language residual was

$$v_{s,\ell} = \tilde{h}_{s,\ell} - u_{s,\ell}, \tag{2}$$

and the orthogonalized language-specific residual was

$$v^{\perp}_{s,\ell} = v_{s,\ell} - \frac{u^{\top}_{s,\ell} v_{s,\ell}}{\|u_{s,\ell}\|_2^2 + \varepsilon} u_{s,\ell}. \tag{3}$$

The two primary conditions were therefore SHARED $= u$ and SPECIFIC $= v^{\perp}$.

5.4. Encoding models and ROI targets

Sentence-level features were placed on each participant's own timeline using the sequel's fixed sentence-span timing operators, convolved with a canonical hemodynamic response function, and sampled at the fMRI TR. The neural targets were ROI-average BOLD time series from atlas-derived semantic, auditory, and control ROI families. The confirmatory inferential unit was always the subject, not the individual time point.

Encoding models used nested leave-one-run-out ridge regression over the nine canonical runs. All nuisance regression, feature standardization, and dimensionality reduction were fit on training data only. Subject-level held-out performance was summarized by Fisher-z transformed Pearson correlations averaged across outer folds. For every block, state, ROI, and language, the primary score of interest was the mean subject-level difference

$$\Delta z = z_{\text{SHARED}} - z_{\text{SPECIFIC}}. \tag{4}$$

5.5. Primary state-family tests

The pre-specified semantic hypothesis compared FFN-side states (FFN and OUTPUT) against attention-side states (ATTN and POST-ATTN) within the semantic ROI family. The pre-specified auditory hypothesis compared the same state families in the opposite direction within the auditory ROI family. Primary statistics were computed with subject-level sign-flip permutation tests (10,000 permutations) and Holm correction across 12 model-language hypotheses. Bootstrap confidence intervals also used 10,000 resamples.

Because the semantic hypothesis failed uniformly in the opposite direction, we added an explicitly labeled exploratory reverse-direction analysis using the subject-level H1 effect table from the main analysis outputs. Those post hoc semantic tests used one-sided sign-flip tests for attention-side > FFN-side effects and Holm correction across the six semantic model-language cells.

5.6. Representative-state selection, token attribution, and deletion validation

Token-level analyses were restricted to representative settings chosen from the state sweep by maximizing the mean subject-level SHAREDminus SPECIFIC difference within each model-language-ROI family cell. This yielded 10 token-scope settings (6 semantic, 4 auditory). Word-level attribution scores were then computed within those representative settings and aggregated into four mutually exclusive classes: content, function, punctuation, and edge tokens.

Deletion validation compared the effect of removing top-attributed words against matched-random deletion within the same sentence, using deletion depths $k = 1$ and $k = 2$. The final summary table contained 20 rows from this stage. Positive values therefore mean that deleting top-attributed words reduced held-out prediction more than deleting matched-random words.

Data, code, and ethics statements

The study is a secondary analysis of a public dataset, LPPC-fMRI [3]. No new human data were collected for this work. Code, manuscript-facing figures, and summary tables are provided with the accompanying materials. The broader internal computation workspace and large intermediate caches are not reproduced in full here.

Competing interests

The author declares no competing interests.

Acknowledgments

Ali Uyar conducted this work as an independent researcher.

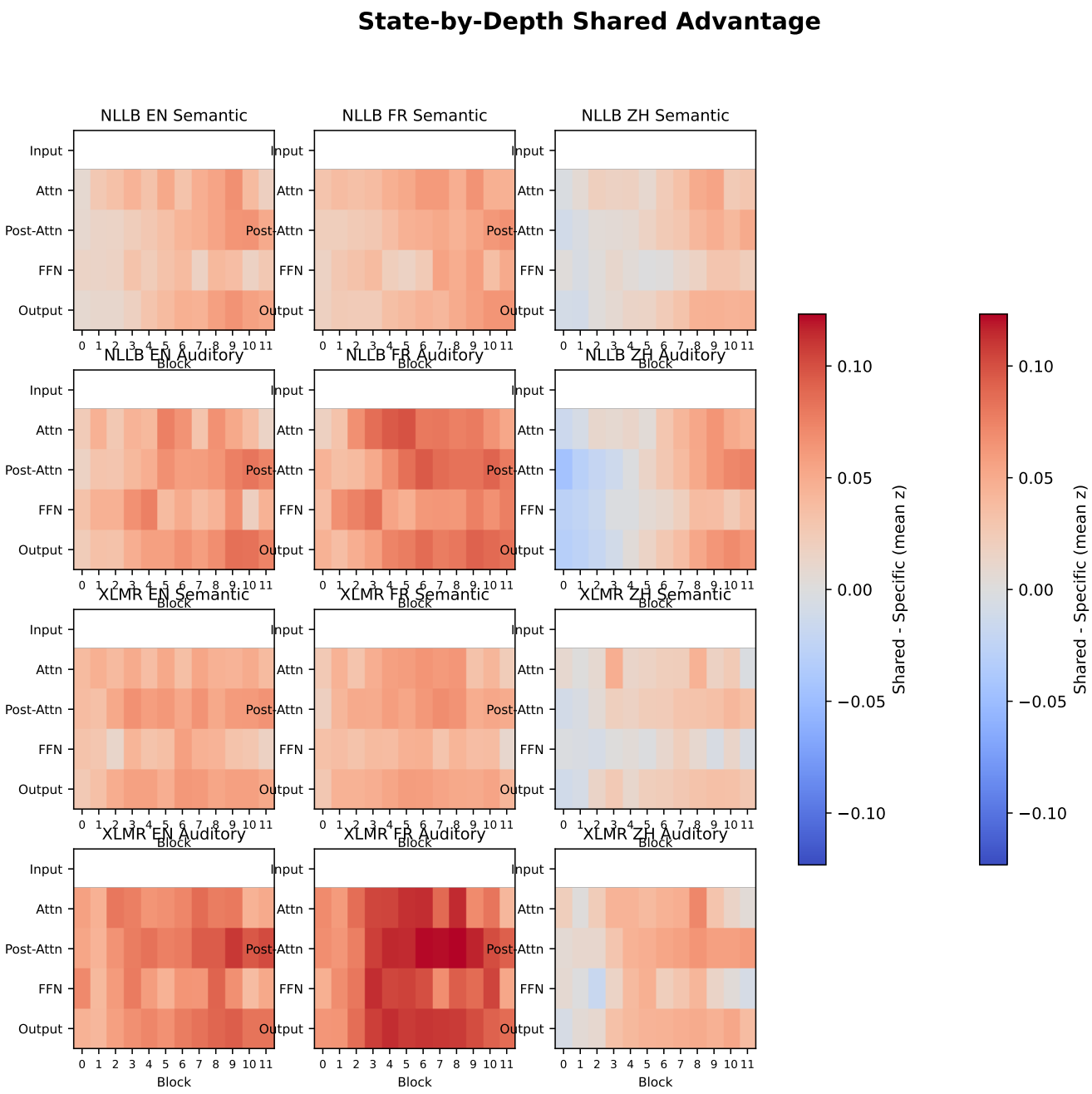

Figure 1: State-by-depth heatmaps of the mean SHAREDminus SPECIFICadvantage. Panels are organized by model, language, and ROI family. The primary mechanistic observation is that the shared advantage is distributed across internal computations rather than confined to a single returned hidden state, and that its strongest bands are predominantly attention-linked rather than FFN-dominated.

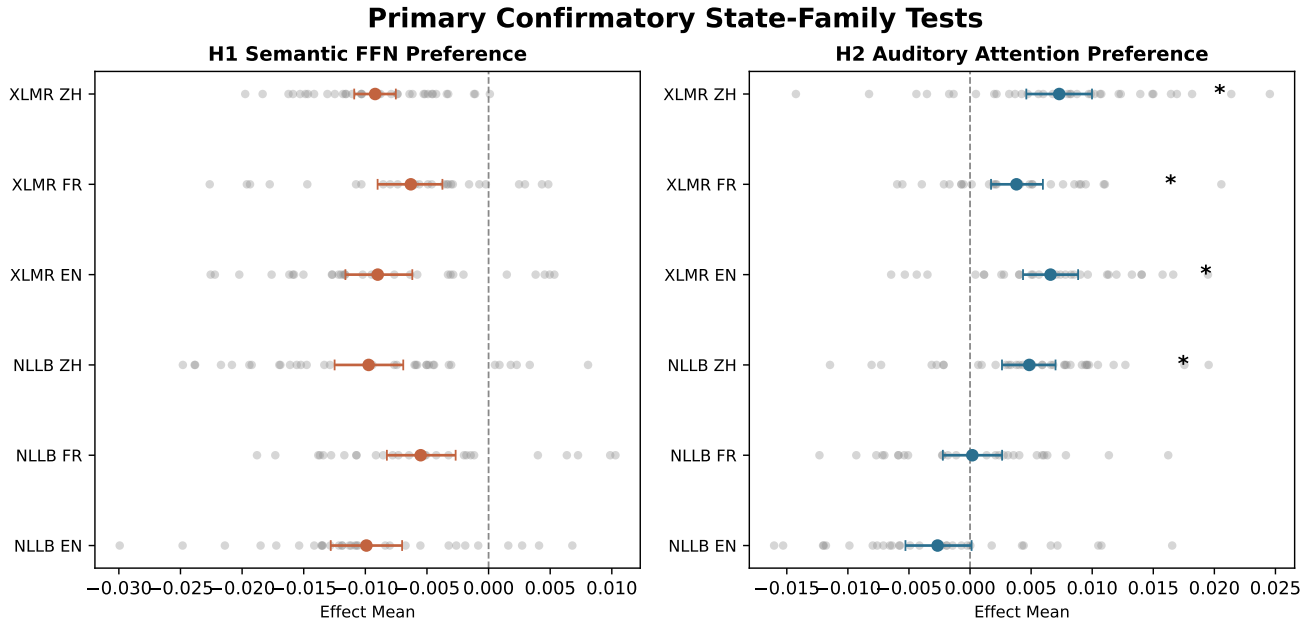

Figure 2: Primary state-family tests. Left: the planned semantic FFN-preference hypothesis (H1) is negative in all six model-language cells, indicating a consistent shift away from FFN-side dominance. Right: the auditory attention-preference hypothesis (H2) is positive in five of six cells and Holm-significant in four of six. Error bars show bootstrap confidence intervals; points show model-language means.

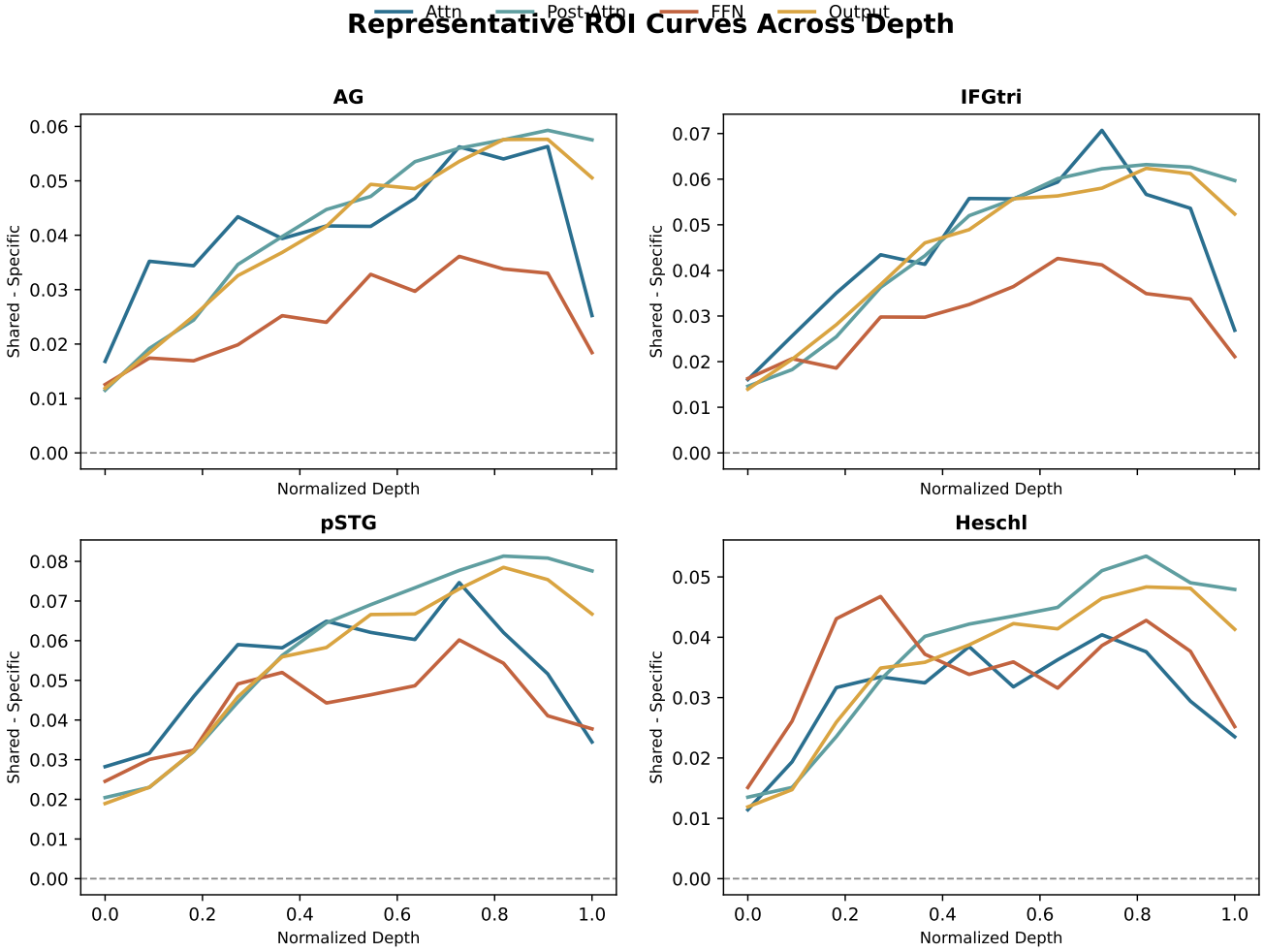

Figure 3: Representative ROI curves across normalized block depth for AG, IFGtri, pSTG, and Heschl. The curves reinforce the family-level result: positive SHARED minus SPECIFIC effects are widespread, but their maxima are largely carried by attention-side or post-attention computations rather than by FFN states.

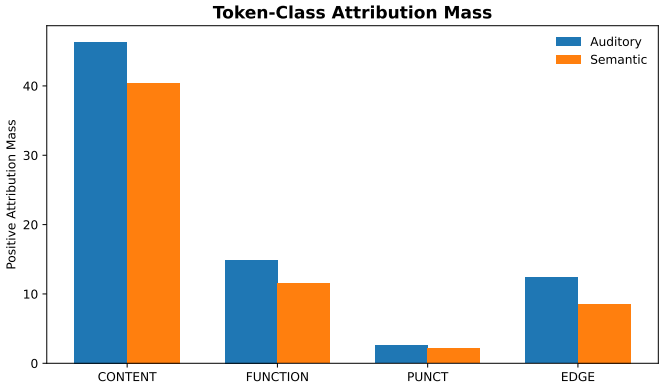

Figure 4: Token-class attribution mass for representative SHARED settings. Semantic settings show a descriptively larger content bias than auditory settings, consistent with a more content-sensitive contribution to the shared brain-predictive signal. Because the representative states were selected from the completed sweep, this token-class contrast is treated as secondary rather than confirmatory evidence.

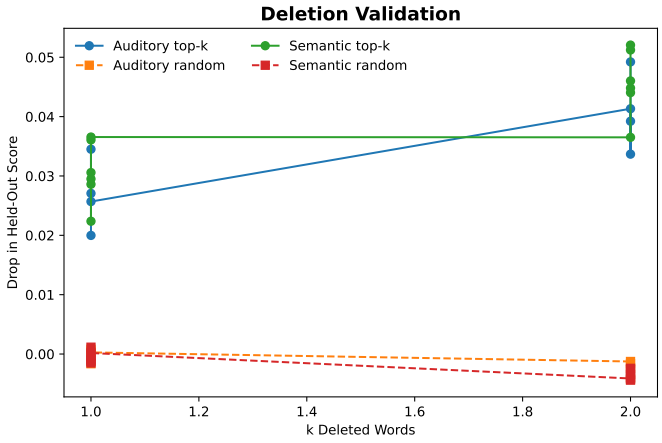

Figure 5: Deletion validation for representative settings. For both ROI families, removing top-attributed words produced a larger drop in held-out performance than matched-random deletion. The effect was positive in all 20 summary rows and increased from $k = 1$ to $k = 2$.

Appendix A. Supplementary figure and table guide

The supplementary materials collect manuscript-facing artifacts that support, but do not define, the headline claim. These include sample and model inventory tables, full confirmatory statistics, representative ROI summaries, full token and deletion tables, a post hoc reverse-direction semantic test summary, and descriptive figures illustrating token examples and ROI preference maps.

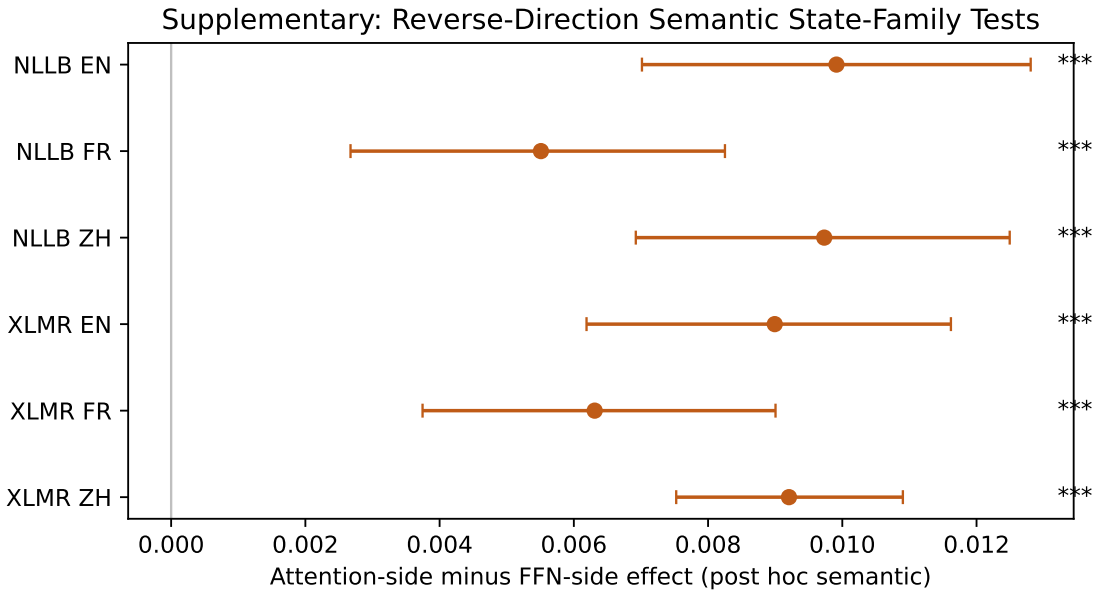

Figure A.6: Exploratory reverse-direction semantic tests derived from the subject-level H1 effects. All six semantic model-language cells favored attention-side over FFN-side states after Holm correction. This figure is supplementary because the tests were performed post hoc after the planned FFN-direction hypothesis failed uniformly.

Appendix B. Supplementary tables

Table B.1: Sample summary.

Language	*n* subjects	Runs	Mean run duration (s)	TR (s)
EN	31	9	625.8	2.0
FR	28	9	652.7	2.0
ZH	35	9	661.6	2.0

Table B.2: Model and state inventory.

Model	Hidden size	Blocks	OUTPUT parity	States
NLLB	1024	12	PASS	ATTN, FFN, OUTPUT, POST-ATTN
XLMR	768	12	PASS	ATTN, FFN, OUTPUT, POST-ATTN

Table B.3: Primary confirmatory state-family statistics from the completed mechanistic sweep. Negative H1 values favor attention-side over FFN-side states; positive H2 values favor attention-side over FFN-side states.

Model	Lang.	Hypothesis	Effect	95% CI low	95% CI high	Holm *p*
NLLB	EN	H1 semantic FFN preference	-0.0099	-0.0128	-0.0070	1

Model	Lang.	Hypothesis	Effect	95% CI low	95% CI high	Holm p
NLLB	EN	H2 auditory attention preference	-0.0027	-0.0053	0.0001	1
NLLB	FR	H1 semantic FFN preference	-0.0055	-0.0083	-0.0027	1
NLLB	FR	H2 auditory attention preference	0.0002	-0.0022	0.0026	1
NLLB	ZH	H1 semantic FFN preference	-0.0097	-0.0125	-0.0069	1
NLLB	ZH	H2 auditory attention preference	0.0048	0.0026	0.0070	0.005
XLMR	EN	H1 semantic FFN preference	-0.0090	-0.0116	-0.0062	1
XLMR	EN	H2 auditory attention preference	0.0066	0.0043	0.0088	0.0012
XLMR	FR	H1 semantic FFN preference	-0.0063	-0.0090	-0.0037	1
XLMR	FR	H2 auditory attention preference	0.0038	0.0017	0.0060	0.007199
XLMR	ZH	H1 semantic FFN preference	-0.0092	-0.0109	-0.0075	1
XLMR	ZH	H2 auditory attention preference	0.0073	0.0046	0.0100	0.0022

Table B.4: Exploratory reverse-direction semantic tests derived from the subject-level H1 effects. These post hoc tests evaluate attention-side minus FFN-side effects after the planned FFN-direction hypothesis failed uniformly.

Model	Lang.	Attention-FFN effect	95% CI low	95% CI high	Holm p
NLLB	EN	0.0099	0.0070	0.0128	3e-05
NLLB	ZH	0.0097	0.0069	0.0125	3e-05
XLMR	EN	0.0090	0.0062	0.0116	3e-05
XLMR	ZH	0.0092	0.0075	0.0109	3e-05
XLMR	FR	0.0063	0.0037	0.0090	3e-05
NLLB	FR	0.0055	0.0027	0.0083	0.00065

Table B.5: Representative ROI summaries. For each model, language, and ROI family, the representative block-state pair maximized the mean subject-level SHARED-SPECIFIC difference.

Model	Lang.	Family	Block	State	ROI	Shared z	Specific z	Δz
NLLB	EN	Auditory	10	OUTPUT	Heschl	0.1177	0.0662	0.0515
NLLB	EN	Auditory	10	OUTPUT	pSTG	0.1586	0.0788	0.0798
NLLB	EN	Semantic	8	ATTN	AG	0.1144	0.0343	0.0801
NLLB	EN	Semantic	8	ATTN	IFGtri	0.1178	0.0366	0.0811
NLLB	FR	Auditory	10	OUTPUT	Heschl	0.1097	0.0513	0.0584
NLLB	FR	Auditory	10	OUTPUT	pSTG	0.1553	0.0666	0.0888
NLLB	FR	Semantic	11	OUTPUT	AG	0.0954	0.0305	0.0650
NLLB	FR	Semantic	11	OUTPUT	IFGtri	0.1041	0.0299	0.0742
NLLB	ZH	Auditory	10	POST-ATTN	Heschl	0.0521	0.0213	0.0309
NLLB	ZH	Auditory	10	POST-ATTN	pSTG	0.1394	0.0704	0.0690
NLLB	ZH	Semantic	9	ATTN	AG	0.0786	0.0391	0.0395
NLLB	ZH	Semantic	9	ATTN	IFGtri	0.1013	0.0415	0.0598
XLMR	EN	Auditory	9	POST-ATTN	Heschl	0.1235	0.0329	0.0906
XLMR	EN	Auditory	9	POST-ATTN	pSTG	0.1631	0.0618	0.1013
XLMR	EN	Semantic	9	POST-ATTN	AG	0.1161	0.0399	0.0762
XLMR	EN	Semantic	9	POST-ATTN	IFGtri	0.1199	0.0460	0.0739
XLMR	FR	Auditory	7	POST-ATTN	Heschl	0.1253	0.0488	0.0764
XLMR	FR	Auditory	7	POST-ATTN	pSTG	0.1671	0.0500	0.1171
XLMR	FR	Semantic	8	ATTN	AG	0.0781	0.0195	0.0586

Model	Lang.	Family	Block	State	ROI	Shared z	Specific z	Δz
XLMR	FR	Semantic	8	ATTN	IFGtri	0.0997	0.0181	0.0816
XLMR	ZH	Auditory	8	ATTN	Heschl	0.0485	0.0219	0.0266
XLMR	ZH	Auditory	8	ATTN	pSTG	0.1439	0.0745	0.0694
XLMR	ZH	Semantic	8	ATTN	AG	0.0721	0.0229	0.0492
XLMR	ZH	Semantic	8	ATTN	IFGtri	0.0908	0.0384	0.0524

Table B.6: Token-class attribution summaries for representative SHARED settings. Positive-mass values are reported in the units used by the generated bundle. Content-bias is descriptive.

Model	Lang.	Family	Condition	Class	Positive mass	Content bias
NLLB	EN	Semantic	SHARED	Content	32.520	0.754
NLLB	EN	Semantic	SHARED	Edge	5.142	0.754
NLLB	EN	Semantic	SHARED	Function	4.284	0.754
NLLB	EN	Semantic	SHARED	Punct	1.184	0.754
NLLB	FR	Semantic	SHARED	Content	27.862	0.733
NLLB	FR	Semantic	SHARED	Edge	5.468	0.733
NLLB	FR	Semantic	SHARED	Function	4.671	0.733
NLLB	ZH	Auditory	SHARED	Content	31.818	0.556
NLLB	ZH	Auditory	SHARED	Edge	9.885	0.556
NLLB	ZH	Auditory	SHARED	Function	13.538	0.556
NLLB	ZH	Auditory	SHARED	Punct	2.016	0.556
NLLB	ZH	Semantic	SHARED	Content	43.480	0.573
NLLB	ZH	Semantic	SHARED	Edge	11.444	0.573
NLLB	ZH	Semantic	SHARED	Function	18.564	0.573
NLLB	ZH	Semantic	SHARED	Punct	2.349	0.573
XLMR	EN	Auditory	SHARED	Content	60.964	0.689
XLMR	EN	Auditory	SHARED	Edge	15.337	0.689
XLMR	EN	Auditory	SHARED	Function	9.318	0.689
XLMR	EN	Auditory	SHARED	Punct	2.901	0.689
XLMR	EN	Semantic	SHARED	Content	46.899	0.705
XLMR	EN	Semantic	SHARED	Edge	9.819	0.705
XLMR	EN	Semantic	SHARED	Function	7.411	0.705
XLMR	EN	Semantic	SHARED	Punct	2.429	0.705
XLMR	FR	Auditory	SHARED	Content	44.217	0.688
XLMR	FR	Auditory	SHARED	Edge	10.363	0.688
XLMR	FR	Auditory	SHARED	Function	9.663	0.688
XLMR	FR	Semantic	SHARED	Content	53.138	0.708
XLMR	FR	Semantic	SHARED	Edge	9.553	0.708
XLMR	FR	Semantic	SHARED	Function	12.366	0.708
XLMR	ZH	Auditory	SHARED	Content	48.458	0.526
XLMR	ZH	Auditory	SHARED	Edge	13.802	0.526
XLMR	ZH	Auditory	SHARED	Function	26.998	0.526
XLMR	ZH	Auditory	SHARED	Punct	2.835	0.526
XLMR	ZH	Semantic	SHARED	Content	38.693	0.531
XLMR	ZH	Semantic	SHARED	Edge	9.761	0.531
XLMR	ZH	Semantic	SHARED	Function	21.819	0.531
XLMR	ZH	Semantic	SHARED	Punct	2.529	0.531

Figure A.7: Illustrative multilingual token-attribution examples for one semantic and one auditory representative setting. Shaded words carry larger positive attribution to the SHARED brain-predictive signal. The examples are shown only for interpretive context; aggregate token-class summaries and deletion validation provide the stronger evidence.

Table B.7: Deletion-validation results by representative setting. Positive differences indicate that removing top-attributed words reduced held-out performance more than matched-random deletion.

Model	Lang.	Family	k	Top-k effect	Random effect	Difference	95% CI low	95% CI high
NLLB	EN	Semantic	1	0.0286	-0.0013	0.0299	0.0246	0.0389
NLLB	EN	Semantic	2	0.0440	-0.0039	0.0480	0.0387	0.0614
NLLB	FR	Semantic	1	0.0224	-0.0015	0.0239	0.0171	0.0342
NLLB	FR	Semantic	2	0.0365	-0.0041	0.0406	0.0289	0.0568
NLLB	ZH	Auditory	1	0.0200	-0.0016	0.0216	0.0161	0.0271
NLLB	ZH	Auditory	2	0.0337	-0.0043	0.0380	0.0285	0.0478
NLLB	ZH	Semantic	1	0.0295	-0.0010	0.0305	0.0216	0.0411
NLLB	ZH	Semantic	2	0.0448	-0.0044	0.0492	0.0368	0.0652
XLMR	EN	Auditory	1	0.0271	0.0005	0.0266	0.0183	0.0361
XLMR	EN	Auditory	2	0.0413	-0.0013	0.0426	0.0308	0.0545
XLMR	EN	Semantic	1	0.0306	-0.0003	0.0309	0.0232	0.0418
XLMR	EN	Semantic	2	0.0460	-0.0034	0.0494	0.0372	0.0671
XLMR	FR	Auditory	1	0.0257	0.0003	0.0254	0.0176	0.0392
XLMR	FR	Auditory	2	0.0392	-0.0016	0.0409	0.0292	0.0625
XLMR	FR	Semantic	1	0.0366	0.0002	0.0364	0.0254	0.0523
XLMR	FR	Semantic	2	0.0521	-0.0024	0.0545	0.0386	0.0751
XLMR	ZH	Auditory	1	0.0345	0.0008	0.0337	0.0273	0.0448
XLMR	ZH	Auditory	2	0.0492	-0.0024	0.0516	0.0396	0.0710
XLMR	ZH	Semantic	1	0.0361	0.0012	0.0349	0.0250	0.0484
XLMR	ZH	Semantic	2	0.0512	-0.0027	0.0539	0.0367	0.0736

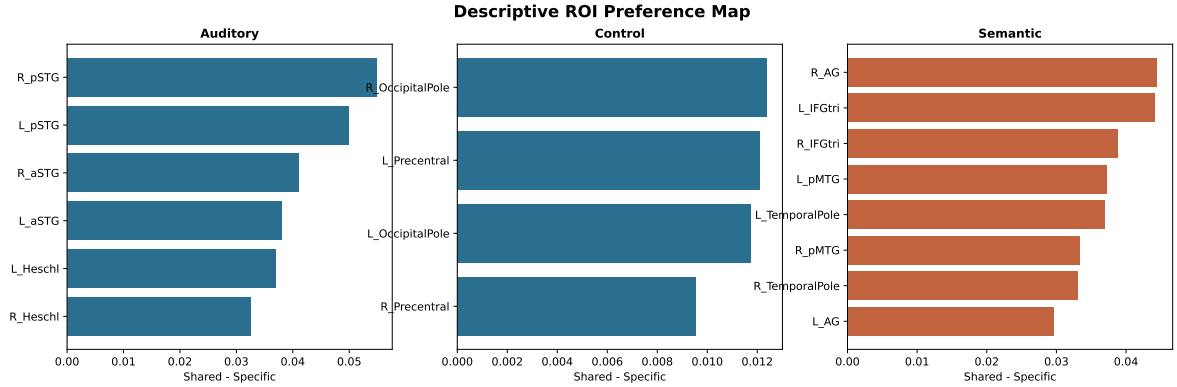

Figure A.8: Descriptive ROI preference map derived from the completed analyses. Values summarize the SHAREDminus SPECIFIC advantage for the representative ROI set and provide additional anatomical context for the main text figures.

References

[1] S. Sonkusare, M. Breakspear, C. S. Guo, Naturalistic stimuli in neuroscience: Critically acclaimed, Trends in Cognitive Sciences 23 (8) (2019) 699–714. doi:10.1016/j.tics.2019.05.004.

[2] E. Simony, C. Chang, Analysis of stimulus-induced brain dynamics during naturalistic paradigms, NeuroImage 216 (2020) 116461. doi:10.1016/j.neuroimage.2019.116461.

[3] J. Li, J. Hale, C. Pallier, Le petit prince multilingual naturalistic fmri corpus, Scientific Data 9 (2022) 530. doi:10.1038/s41597-022-01625-7.

[4] L. Wehbe, B. Murphy, P. Talukdar, A. Fyshe, A. Ramdas, T. Mitchell, Simultaneously uncovering the patterns of brain regions involved in different story reading subprocesses, PLoS ONE 9 (11) (2014) e112575. doi:10.1371/journal.pone.0112575.

[5] M. Toneva, L. Wehbe, Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain), in: Advances in Neural Information Processing Systems, Vol. 32, 2019.

[6] C. Caucheteux, J.-R. King, Brains and algorithms partially converge in natural language processing, Communications Biology 5 (2022) 134. doi:10.1038/s42003-022-03036-1.

[7] M. Schrimpf, I. A. Blank, G. Tuckute, C. Kauf, E. A. Hosseini, N. Kanwisher, J. B. Tenenbaum, E. Fedorenko, The neural architecture of language: Integrative modeling converges on predictive processing, Proceedings of the National Academy of Sciences 118 (45) (2021) e2105646118. doi:10.1073/pnas.2105646118.

[8] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, in: Proceedings of ACL, 2020, pp. 8440–8451.

[9] NLLB Team, M. R. Costa-jussà, J. Cross, O. Çelebi, M. Elbayad, K. Heafield, et al., No language left behind: Scaling human-centered machine translation, arXiv preprint arXiv:2207.04672 (2022).