Brain-Grounded Axes for Reading and Steering LLM States

Sandro Andric
sandro.andric@nyu.edu

Abstract

Interpretability methods for large language models (LLMs) typically derive directions from textual supervision, which can lack external grounding. We propose using human brain activity not as a training signal but as a coordinate system for reading and steering LLM states. Using the SMN4Lang MEG dataset, we construct a word-level brain atlas of phase-locking value (PLV) patterns and extract latent axes via ICA. We validate axes with independent lexica and NER-based labels (POS/log-frequency used as sanity checks), then train lightweight adapters that map LLM hidden states to these brain axes without fine-tuning the LLM. Steering along the resulting brain-derived directions yields a robust lexical (frequency-linked) axis in a mid TinyLlama layer, surviving perplexity-matched controls, and a brain-vs-text probe comparison shows larger log-frequency shifts (relative to the text probe) with lower perplexity for the brain axis. A function/content axis (axis 13) shows consistent steering in TinyLlama, Qwen2-0.5B, and GPT-2, with PPL-matched text-level corroboration. Layer-4 effects in TinyLlama are large but inconsistent, so we treat them as secondary (Appendix). Axis structure is stable when the atlas is rebuilt without GPT embedding-change features or with word2vec embeddings ($|r| = 0.64$–0.95 across matched axes), reducing circularity concerns. Exploratory fMRI anchoring suggests potential alignment for embedding change and log frequency, but effects are sensitive to hemodynamic modeling assumptions and are treated as population-level evidence only. These results support a new interface: neurophysiology-grounded axes provide interpretable and controllable handles for LLM behavior.

1 Introduction

Most LLM interpretability discovers directions using text-derived labels or optimization objectives. Such directions can be effective yet are rarely grounded in external measurements of human cognition. We ask whether brain activity can define a stable, externally grounded coordinate system that enables both reading and steering of LLM representations. Our key idea is to compute a brain-derived atlas of semantic activity and use its axes as the target space for LLM projections. We do not optimize the LLM; instead we train a lightweight adapter and test whether steering along those axes produces reliable, interpretable shifts in generated text.

Contributions: - We build a MEG-derived word-level brain atlas of PLV connectivity and extract latent axes with stability/robustness checks. - We show that a lightweight adapter can map LLM hidden states into this brain axis space without fine-tuning. - We demonstrate robust steering for a frequency-linked axis (TinyLlama L11) and cross-model steering for a function/content axis (GPT-2, Qwen-0.5B, TinyLlama), including PPL-matched and brain-vs-text probe controls.

Preprint.

2 Related Work

Prior work aligns neural data with language models for encoding/decoding and representation comparisons [15, 8, 4, 2], uses brain data to interpret NLP models [17], and demonstrates semantic reconstruction from brain recordings [16]. Recent steering methods use activation-addition and contrastive vectors [19, 13]. Our contribution differs in two ways: (1) we derive axes from MEG connectivity geometry rather than textual supervision, and (2) we treat brain axes as a fixed interface for reading and steering, not as a reward for optimization.

3 Dataset

We use SMN4Lang (OpenNeuro ds004078 v1.2.1) [20, 21], a synchronized MEG/fMRI dataset of naturalistic story listening. We use preprocessed MEG sensor-level data and word-level timing annotations. The dataset includes 12 subjects and 60 stories.

4 Methods

4.1 MEG PLV Atlas

We compute phase-locking value (PLV) in theta band (4–8 Hz) over sliding windows (2.0 s length, 0.5 s step) on gradiometer channels [10]. For each window, we compute PLV connectivity and compress the edge vector using PCA to 128 dimensions (edge-PCA). Filtering and analytic phase extraction are implemented with MNE-Python [5]. This yields a time-indexed PLV state vector per run.

4.2 Word-Level Brain Atlas

We align PLV windows to word-level timing annotations and build a word-level atlas using ridge regression [6] with features: embedding change (GPT), word log-frequency, and POS id. Lags of 0.0, 0.5, and 1.0 s are included. The model predicts PLV-PCA states per window; each window is assigned to the most recent word onset, and we average predicted PLV vectors by word type to obtain a word-level atlas per subject. We use predicted PLV states (rather than single-trial PLV averages) to improve signal-to-noise and stabilize word-level estimates. Subject-level atlases are then averaged across subjects to form a cross-subject atlas. To test circularity, we rebuild the atlas with embedding change removed (logfreq+POS only) and with word2vec embeddings [11], then align axes by correlation. Atlas predictions are generated out-of-fold by run (5-fold): for each fold, models are trained on a subset of runs and used to predict held-out runs, and the word atlas is built from these OOF predictions (full-data weights are saved for reference).

4.3 Axis Discovery

We apply ICA to the averaged word atlas and obtain latent axes [7]. We fix the number of components to 20 a priori for stability across runs. We validate axes using independent label sources:

- CKIP POS/NER for function/content, noun/verb, and animate labels [9] (temporal heuristics used only for exploratory checks)
- MELD-SCH concreteness [18, 22] and Chinese valence/arousal lexica [24]

Permutation tests confirm axis-label alignment. Axis labels in the paper (e.g., function/content, animacy) are assigned by the strongest atlas-label association and are not selected using steering outcomes. Axis IDs refer to fixed ICA component indices; when comparing across layers or models, we align axis direction by correlation and report signed effects after alignment (orientation is arbitrary).

4.4 LLM Adapter

We extract TinyLlama word-level hidden states for all stories [25] and map them to brain axes using ridge regression (standardized, alpha chosen by CV). This yields an adapter $f(h) = Wh + b$ that predicts axis scores per word.

4.5 Steering

We steer TinyLlama by adding a normalized axis vector to hidden states at selected layers during generation. Concretely, for axis k we use the corresponding adapter weight vector W_k

(scaled back if standardization is used), L2-normalize it, and add $\alpha W_k/\|W_k\|$ to the hidden state at the chosen layer for all positions in each forward pass (prompt and generated tokens). We test strengths $\{-5, -2, -1, 0, 1, 2, 5\}$ over 50 prompts with 4 samples per strength (256 tokens each) using batched generation, and evaluate:

- Adapter-score shift (pos vs neg, t-test + permutation)
- Strength correlation
- Fluency (perplexity) shift
- Text-level POS/NER metrics

We correct for multiple comparisons across layers and axes with FDR [1].

5 Results

5.1 Axis Validation

Word-level validation confirms strong axis-label alignment on non-POS lexica:

- Animate vs inanimate (confound-controlled): axis 2, residualized $d = 0.53$ (logfreq+surprisal+length), matched $d = 0.70$ with CI [0.53, 0.86] (n=330)
- Lexical frequency: axis 15, $r = 0.51$ with logfreq (n=13,632)

Lexica validation: concreteness (axis 2, $r = -0.128$, $p = 0.001$), valence (axis 15, $r = 0.114$, $p = 0.001$), arousal (axis 15, $r = 0.084$, $p = 0.001$). Axis 15 is strongly frequency-linked, so valence/arousal effects are treated as secondary. POS-derived validations (function/content, noun/verb) are reported as sanity checks only (Appendix).

Confound control shows that animacy is not driven by logfreq/surprisal/length: axis 2 remains significant after residualization ($d = 0.53$, $p = 7.1 \times 10^{-5}$), while axis 15 attenuates (residualized $d = 0.13$, $p = 0.25$).

Out-of-fold (OOF) atlas predictions preserve the axis geometry: the best-match correlations for axes 13/19/15/2 are $|r| = 0.82$–0.97 relative to the base atlas. For reproducibility we match OOF axes to the original indices by correlation (2→3, 13→11, 15→5, 19→12) and keep the base labels throughout the paper.

Leakage-robust control: in a wPLI run (4 subjects, runs 10–20, decim=5), the lexical-frequency axis remains robust ($|r| = 0.744$), while other axes show weaker correspondence ($|r| \approx 0.35$–0.48), suggesting non-lexical axes are less stable under leakage-robust connectivity at this reduced-data setting.

5.2 Embedding-Change Ablations

Rebuilding the atlas without GPT embedding-change features (logfreq+POS only) or with word2vec embeddings yields axes highly correlated with the base atlas ($|r| = 0.82$–0.95 for axes 13/19/15/2 under no-embedding, $|r| = 0.64$–0.93 under word2vec), indicating the axis geometry is not driven by GPT features. Dropping POS features substantially alters the POS-linked axes ($|r| < 0.3$), confirming these labels are not independent; we therefore do not treat POS-derived axis validation as primary evidence. Dropping log-frequency features markedly reduces lexical-axis alignment (axis 15 correlation drops to $|r| = 0.19$), so we treat this axis as supervised rather than emergent.

5.3 Axis Reliability and Cross-Subject Generalization

We ran a rigorous reliability analysis with bootstrap confidence intervals and permutation tests across 60 axis-lexicon pairs. After FDR correction, 25/60 tests remain significant (expected 3 by chance). The strongest effects include concreteness (axis 2, $r = -0.128$ with CI [-0.162, -0.092]), valence (axis 15, $r = 0.114$ with CI [0.082, 0.141]), and arousal (axis 15, $r = 0.084$ with CI [0.056, 0.114]). Partial correlations controlling for word length show negligible change for concreteness; axis 15 remains strongly tied to log frequency ($r = 0.51$).

Cross-subject validation (odd vs even splits) shows 12 axis-dimension pairs replicating with $|r| > 0.05$. We align axes by correlation and report absolute values to avoid sign ambiguity (ICA orientation is arbitrary). Examples include the matched concreteness axis

(test $|r| = 0.079 \pm 0.014$), valence axis (test $|r| = 0.093 \pm 0.023$), and arousal axis (test $|r| = 0.083 \pm 0.005$). This supports stable semantic structure across subjects.

5.4 Adapter Transfer to LLM

The TinyLlama adapter generalizes to held-out words (9,149 matched words):

- Axis 13: $r = 0.238$ ($p < 10^{-24}$)
- Axis 19: $r = 0.499$ ($p < 10^{-115}$)
- Axis 15: $r = 0.461$ ($p < 10^{-96}$)
- Axis 2: $r = 0.624$ ($p < 10^{-197}$)

Cross-model transfer (Qwen2-0.5B). Using the same atlas, a lightweight adapter trained on Qwen2-0.5B hidden states [23] also predicts brain axes on held-out words (9,149 matched): axis 13 $r = 0.264$, axis 19 $r = 0.528$, axis 15 $r = 0.435$, axis 2 $r = 0.637$ (all $p \ll 10^{-30}$). This supports the "interface" framing beyond a single LLM. Steering in Qwen2-0.5B shows significant adapter-score shifts at layer 12 for axes 13/15/2 (perm $p \leq 0.002$), with axis 19 weaker (perm $p = 0.029$), and a strong axis 13 effect at layer 4. GPT-2 [14] shows a consistent axis 13 effect across layers (best L6 $d = 0.30$, perm $p = 0.001$), while axis 15 is null in GPT-2. These results provide cross-model steering evidence for axis 13; axis 15 appears model-dependent (present in TinyLlama and Qwen-0.5B, absent in GPT-2).

5.5 Steering (Primary Evidence)

Primary result: a lexical frequency-linked (supervised) axis (axis 15) at TinyLlama layer 11 shows robust steering with PPL-matched confirmation and a brain-vs-text probe efficiency advantage. We treat independent text-level metrics as primary evidence and use adapter-score shifts only as a manipulation check. Layers tested: 4, 11, 20 (TinyLlama). In the batched 50-prompt run, axis 15 yields $d = 0.545$ (perm $p = 0.001$). Under PPL-matched controls with the same 50-prompt setup, the effect remains ($d = 0.571$, perm $p = 0.001$). A brain-vs-text probe comparison at the same layer shows that the brain axis produces large log-frequency shifts with lower perplexity ($d = 0.9246$; PPL $d = -0.2183$), while a text-only logfreq probe shifts in the opposite direction and increases PPL ($d = -0.3021$; PPL $d = 0.4549$). We therefore treat log-frequency as the primary independent outcome for axis 15 and use adapter-score shifts as a manipulation check.

To benchmark against standard representation engineering, we computed an Activation Addition (ActAdd) baseline using the mean difference between the top/bottom 50 log-frequency words at TinyLlama layer 11. As expected, this supervised ActAdd vector produced a larger log-frequency shift ($d = 1.6666$, perm $p = 0.001$) [19]. The brain axis remained competitive ($d = 0.9246$) and differed in two key ways: it significantly improved fluency (PPL $d = -0.2183$, perm $p = 0.001$), while ActAdd had no significant PPL effect (PPL $d = -0.0740$, perm $p = 0.237$), and the directions are nearly orthogonal (cosine similarity 0.0104).

Across models, axis 13 (function/content) shows consistent steering: TinyLlama L11 ($|d| = 0.208$, perm $p = 0.001$), Qwen-0.5B L4 ($d = 0.436$, perm $p = 0.001$), and GPT-2 L6 ($d = 0.300$, perm $p = 0.001$). Perplexity-matched controls for axis 13 in TinyLlama L11 and Qwen L4 preserve the effect and yield strong text-level function/content shifts. Other layers and axes show weaker or inconsistent effects; full summaries appear in Appendix (Table 1, Fig. 1).

TinyLlama layer 4 shows large raw shifts but sign flips and lower prompt-level consistency across axes; we therefore keep layer-4 effects as secondary evidence.

model	axis	label	selected layer	d	perm_p
Qwen2-0.5B	13	function_content	4	0.436	0.000999
Qwen2-0.5B	15	lexical_freq	12	0.380	0.000999
Qwen2-0.5B	19	noun_verb	16	0.212	0.000999
Qwen2-0.5B	2	animacy	12	0.181	0.002
GPT-2	13	function_content	6	0.300	0.000999

model	axis	label	selected layer	d	perm_p
GPT-2	15	lexical_freq	6	-0.032	0.57
GPT-2	19	noun_verb	3	0.140	0.018
GPT-2	2	animacy	12	0.109	0.0509
TinyLlama	13	function_content	11	-0.208	0.000999
TinyLlama	15	lexical_freq	11	0.545	0.000999
TinyLlama	19	noun_verb	11	0.316	0.000999
TinyLlama	2	animacy	11	0.194	0.002

Selected layer per model and axis. For TinyLlama we report the primary layer (11) to avoid unstable layer-4 effects.

As a baseline, a matched random-direction steering run at layer 11 (same prompts/strengths/samples/tokens) yields no effect (axis 15 random: $p = 0.88$, $d = -0.02$), indicating that arbitrary directions do not produce the observed shifts.

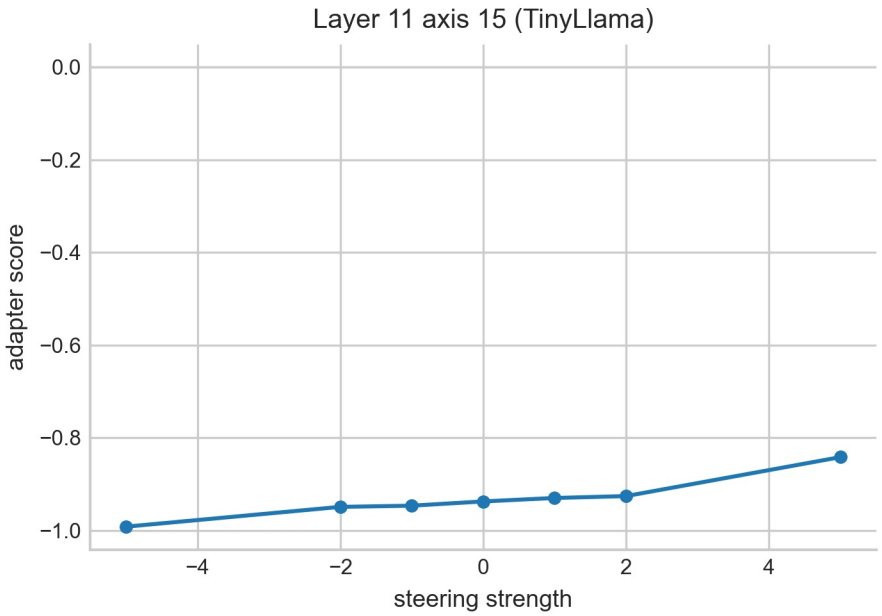

Figure 1: Layer 11 axis 15 steering curve (TinyLlama; batched 50 prompts). Adapter-score means are plotted per strength.

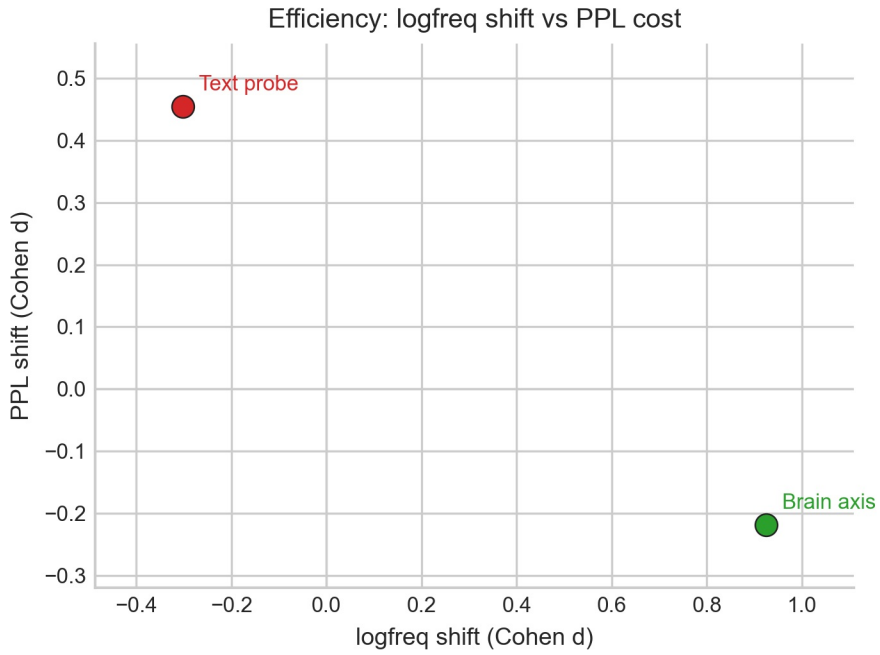

Figure 2: Efficiency comparison for brain axis, text probe, and ActAdd steering (TinyLlama L11). Brain-axis steering yields a large log-frequency shift with improved perplexity; ActAdd yields a larger shift with no significant PPL change.

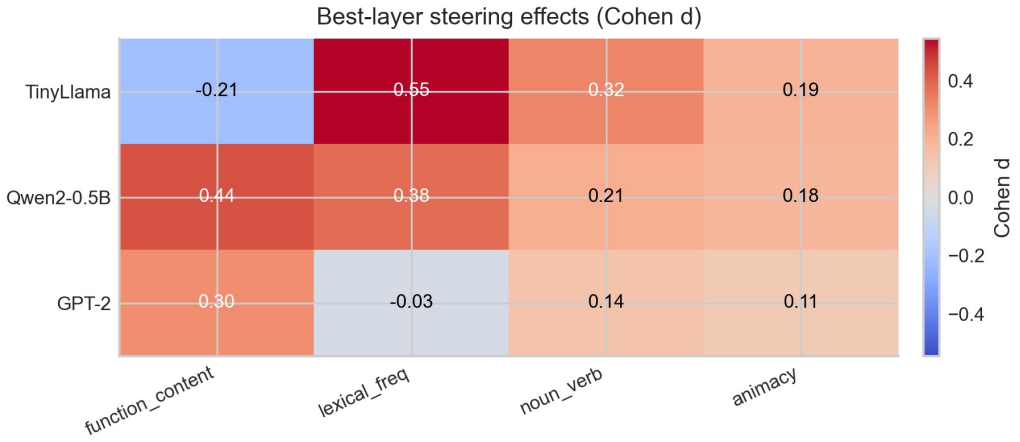

Figure 3: Selected-layer steering effects by model and axis (cell color encodes Cohen d; numbers show d values).

5.6 Perplexity-Matched Controls

The primary effect (layer 11, axis 15) remains significant under perplexity matching (see above). We also ran PPL-matched controls for axis 13 in TinyLlama L11 and Qwen L4; both preserve adapter-score effects and show strong text-level function/content shifts (Appendix).

5.7 External Text-Level Validation

POS/NER metrics show partial external corroboration:

- Log-frequency mean: significant at layer 11 axis 15 (perm $p = 0.001$; $d = 0.765$)

- Function/content ratio: significant at layer 11 axis 13 (perm $p = 0.001$; $d = 0.739$) and under PPL-matched controls in TinyLlama L11 and Qwen L4 (perm $p = 0.001$)
- Animate rate: significant at layer 11 axis 2 (perm $p = 0.001$; $d = 0.169$)
- Noun ratio: significant at layer 11 axis 19 (perm $p = 0.001$; $d = 0.817$) and weaker at layer 4 axis 19 (perm $p = 0.015$; $d = -0.156$)

These text-level shifts are treated as secondary evidence because adapter-score effects are the primary criterion; we therefore report both to avoid measurement-within-the-loop concerns. POS-based text metrics (function/content, noun ratio) are not independent because POS tags are used in atlas construction and are treated as sanity checks only.

5.8 fMRI Anchoring (Exploratory)

We ran an encoding-based fMRI anchoring test (ridge) and varied the HRF. Using a 4 s peak HRF, pooled permutation tests across subjects and runs (n=240) show weak but significant effects for embedding change (observed mean abs corr 0.0888 vs null 0.0718, perm $p = 0.001$) and log frequency (0.0815 vs 0.0725, perm $p = 0.009$). POS id is not significant (perm $p = 0.152$). Per-subject effects are inconsistent, so we treat fMRI anchoring as population-level evidence only; sensitivity analyses are in Appendix.

6 Discussion

Brain-derived axes can steer LLM behavior without LLM fine-tuning. Strongest evidence appears for a lexical (frequency-linked) axis in a mid layer; the animacy axis (axis 2) is robust in the atlas but is not the strongest steering effect. Fluency-controlled analyses confirm the primary effect is not driven by perplexity shifts. External text metrics provide the primary independent validation, while adapter-score shifts are treated as manipulation checks.

A key concern is circularity: do brain-derived axes simply recover text statistics? The brain-vs-text probe comparison argues against this. The brain axis shifts log-frequency in the intended direction with lower perplexity cost ($d = 0.9246$; PPL $d = -0.2183$), while a text-only logfreq probe shifts in the opposite direction and increases perplexity ($d = -0.3021$; PPL $d = 0.4549$). This suggests that mapping through neurophysiology can act as a functional filter, isolating causal structure that a direct text probe does not capture.

Baseline comparisons also reveal distinct geometric mechanisms. The frequency axis (axis 15) is nearly orthogonal to the supervised ActAdd vector (cosine similarity 0.0104), yet both steer log-frequency. This suggests multiple frequency-related subspaces: a direct statistical direction (ActAdd) and a more naturalistic direction derived from the brain atlas. Steering along the brain-derived path significantly improves perplexity, while ActAdd yields no significant PPL change.

7 Limitations

The atlas is sensor-space and may include field spread [12, 3]. Text-level metrics are limited by automatic tagging accuracy; we treat them as primary but imperfect evidence and use adapter-score shifts as manipulation checks alongside PPL controls. POS-derived validations are not independent because POS features are used in atlas construction; we report them as sanity checks only. Some axes show sign reversals across layers, reflecting orientation ambiguity or layer-specific encoding. The strongest steering axis is frequency-linked (supervised), so semantic purity is limited; future work should disentangle lexical confounds. Cross-model transfer is shown for adapter readout (Qwen2-0.5B), and cross-model steering is supported for axis 13 in Qwen2-0.5B and GPT-2, but axis 15 is model-dependent (present in TinyLlama/Qwen-0.5B, absent in GPT-2). fMRI anchoring yields weak global with inconsistent per-subject reliability. We do not claim exclusivity; other axes and layers may be steerable but did not meet our strict robustness criteria here.

8 Ethics

The work uses public, de-identified neuroimaging data. Steering results are modest and should not be interpreted as mind-reading or behavioral prediction of individuals.

9 Reproducibility

Code, configs, and minimal reproduction instructions are available at: https://github.com/sandroandric/Brain-Grounded-Axes-for-Reading-and-Steering-LLM-States.git. The repository documents required inputs and the command-line steps to reproduce all reported results.

10 Supplementary Steering Analyses

Layer	Axis	d	perm p
4	function_content	0.1922	0.001998
4	lexical_freq	-0.8490	0.001
4	noun_verb	-0.2726	0.001
4	animacy	-0.1227	0.03397
11	function_content	-0.2079	0.001
11	lexical_freq	0.5450	0.001
11	noun_verb	0.3163	0.001
11	animacy	0.1936	0.001998
20	function_content	-0.1128	0.04296
20	lexical_freq	0.2580	0.001
20	noun_verb	-0.0106	0.845
20	animacy	0.1340	0.01898

Full steering summary across layers and axes (TinyLlama, batched 50-prompt run). perm p is from permutation tests on adapter-score shift; signs are reported after axis matching (orientation is arbitrary).

10.1 Appendix Baselines

Baseline (TinyLlama L11, 50 prompts)	logfreq d	logfreq p	PPL d	PPL p
Random dir. (axis 15)	-0.02	0.88	n/a	n/a
ActAdd (top/bot 50 logfreq)	1.6666	0.001	-0.0740	0.237
Brain axis (axis 15)	0.9246	0.001	-0.2183	0.001

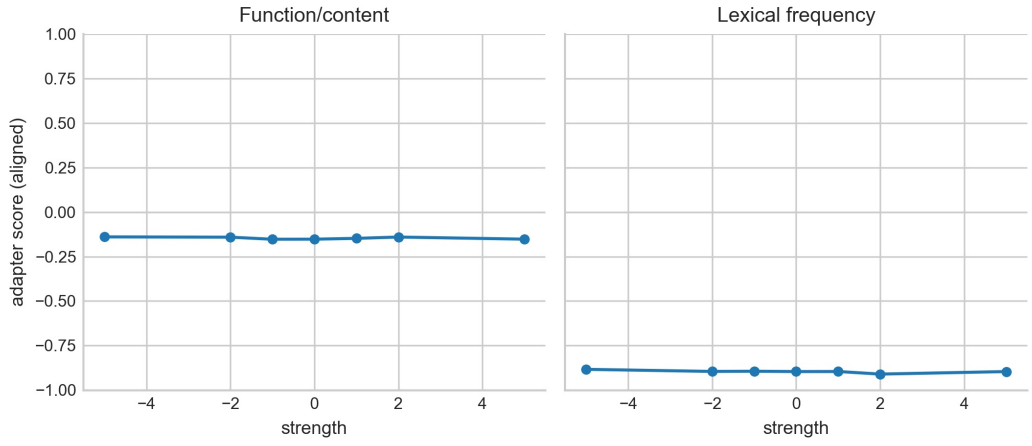

Figure 4: Layer 4 steering effects for the function/content axis (left) and lexical frequency axis (right), aligned to the layer-11 axis orientation. Plots show adapter-score shift across strengths.

Perplexity-matched checks for secondary effects:

- TinyLlama L11, axis 13: adapter shift $d = -0.2355$ (perm $p = 0.001$); function/content ratio $d = 0.6086$ (perm $p = 0.001$).
- Qwen L4, axis 13: adapter shift $d = 0.4364$ (perm $p = 0.001$); function/content ratio $d = 1.0620$ (perm $p = 0.001$).

11 Supplementary fMRI Anchoring Sensitivity

HRF-off anchoring yields significant pooled effects (embedding change $p = 0.001$, logfreq $p = 0.001$, pos_id $p = 0.006$) but with smaller absolute correlations. The default 6 s peak HRF shows only marginal pooled effects (embedding change $p = 0.071$, logfreq $p = 0.082$, pos_id $p = 0.039$). These variants reinforce that fMRI anchoring is weak and sensitive to HRF assumptions.

References

[1] 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*, 57(1):289–300, 1995. doi: 10.1111/j.2517-6161.1995.tb02031.x.

[2] Charlotte Caucheteux and Jean-Remi King. Brains and algorithms partially converge in natural language processing. *Communications Biology*, 5:134, 2022. doi: 10.1038/s42003-022-03036-1.

[3] Gemma L. Colclough, Matthew J. Brookes, Stephen M. Smith, and Mark W. Woolrich. A symmetric multivariate leakage correction for meg connectomes. *NeuroImage*, 117: 439–448, 2015. doi: 10.1016/j.neuroimage.2015.03.071.

[4] Jon Gauthier and Roger Levy. Linking artificial and human neural representations of language. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 529–539, 2019. doi: 10.18653/v1/D19-1050.

[5] Alexandre Gramfort et al. Meg and eeg data analysis with mne-python. *Frontiers in Neuroscience*, 7:267, 2013. doi: 10.3389/fnins.2013.00267.

[6] Arthur E. Hoerl and Robert W. Kennard. Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970. doi: 10.1080/00401706.1970.10488634.

[7] Aapo Hyvarinen and Erkki Oja. Independent component analysis: algorithms and applications. *Neural Networks*, 13(4-5):411–430, 2000. doi: 10.1016/S0893-6080(00)00026-5.

[8] Shailee Jain and Alexander G. Huth. Incorporating context into language encoding models for fmri. In *Advances in Neural Information Processing Systems 31 (NeurIPS 2018)*, pages 6628–6637, 2018.

[9] CKIP Lab. Ckiptagger: Ckip neural chinese word segmentation, pos tagging, and ner, n.d. URL https://ckip.iis.sinica.edu.tw/service/ckiptagger/.

[10] Jean-Philippe Lachaux, Eugenio Rodriguez, Jacques Martinerie, and Francisco J. Varela. Measuring phase synchrony in brain signals. *Human Brain Mapping*, 8(4):194–208, 1999. doi: 10.1002/(SICI)1097-0193(1999)8:4<194::AID-HBM4>3.0.CO;2-C.

[11] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space, 2013. URL https://arxiv.org/abs/1301.3781.

[12] Guido Nolte, Ou Bai, Lewis Wheaton, Zoltan Mari, S. Vorbach, and Mark Hallett. Identifying true brain interaction from eeg data using the imaginary part of coherency. *Clinical Neurophysiology*, 115(10):2292–2307, 2004. doi: 10.1016/j.clinph.2004.04.029.

[13] Nina Panickssery, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. Steering llama 2 via contrastive activation addition, 2023. URL https://arxiv.org/abs/2312.06681.

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

[15] Martin Schrimpf et al. The neural architecture of language: Integrative modeling converges on predictive processing. *PNAS*, 118(45):e2105646118, 2021. doi: 10.1073/pnas.2105646118.

[16] Jerry Tang et al. Semantic reconstruction of continuous language from non-invasive brain recordings. *Nature Neuroscience*, 26:858–866, 2023. doi: 10.1038/s41593-023-01304-9.

[17] Mariya Toneva and Leila Wehbe. Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain). In *Advances in Neural Information Processing Systems (NeurIPS 2019)*, pages 14954–14964, 2019.

[18] Yiu-Kei Tsang et al. Meld-sch: A megastudy of lexical decision in simplified chinese. *Behavior Research Methods*, 50:1763–1777, 2018. doi: 10.3758/s13428-017-0944-0.

[19] Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering, 2023. URL https://arxiv.org/abs/2308.10248.

[20] Shaonan Wang, Xiaohan Zhang, Jiajun Zhang, and Chengqing Zong. A synchronized multimodal neuroimaging dataset for studying brain language processing. *Scientific Data*, 9:590, 2022. doi: 10.1038/s41597-022-01708-5.

[21] Shaonan Wang, Xiaohan Zhang, Jiajun Zhang, and Chengqing Zong. Openneuro ds004078 (smn4lang) v1.2.1, 2023. URL https://doi.org/10.18112/openneuro.ds004078.v1.2.1.

[22] X. Xu and J. Li. Concreteness/abstractness ratings for two-character chinese words in meld-sch. *PLOS ONE*, 15(6):e0232133, 2020. doi: 10.1371/journal.pone.0232133.

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

[24] Liang-Chih Yu, Lung-Hao Lee, Shuai Hao, Jin Wang, Yunchao He, Jun Hu, K. Robert Lai, and Xuejie Zhang. Building chinese affective resources in valence-arousal dimensions. In *Proceedings of NAACL-HLT 2016*, pages 540–545, 2016.

[25] Peiyuan Zhang, Guangtao Zeng, Tianduo Wang, and Wei Lu. Tinyllama: An open-source small language model, 2024. URL https://arxiv.org/abs/2401.02385.