Steering Transformer Attention with Human EEG

Claire Short [*1], **Steven Basart** [*2], **Sinem Erisken** [*3]

[1]MATS
[2]Center for AI Safety
[3]Independent

Abstract

Modern LLMs differ fundamentally from the human brain in architecture and computational mechanisms, yet recent work reveals surprising representational alignments between these systems. Here we test whether noninvasive neural signals can directly steer transformer attention at inference time. Using an InstABoost-style framework, we inject EEG-derived attention weights (suppressing alpha, enhancing theta/gamma bands) into early layers of Llama-3.2-3B without additional training. On reading comprehension tasks from the ZuCo dataset, we observe modest but consistent improvements (0.4-1.4% absolute gain), particularly when using population-averaged EEG from task-specific reading conditions. Control experiments with shuffled or misaligned EEG confirm these gains stem from temporal alignment between neural dynamics and word sequences. While preliminary, these results suggest that human attentional rhythms encode routing information that can productively guide artificial attention mechanisms, opening possibilities for neural-augmented language models.

Code — https://github.com/xksteven/neural_llm_repe/

Introduction

Large language models and human brains process language through vastly different mechanisms—transformers use parallel, feedforward computation with explicit attention, while brains employ recurrent, spike-based processing with distributed representations. Despite these differences, mounting evidence reveals systematic correspondences: intermediate LLM layers predict cortical activity during narrative listening (correlations up to $r = 0.92$), both systems engage in predictive processing, and attention heads map to discrete cortical regions exhibit hierarchical, context-sensitive feature extraction that mirrors cortical processing (Hickok and Poeppel 2007; Lerner et al. 2011; Ding et al. 2017; Sheng et al. 2018; Mischler et al. 2024; Kumar et al. 2023). These alignments raise an intriguing question: can human neural signals directly guide transformer computations?

Our approach is straightforward: we extract frequency-band power (theta, alpha, gamma) from the ZuCo dataset's (Hollenstein et al. 2018, 2019) word-aligned EEG recordings, compute attention multipliers that suppress alpha and enhance theta/gamma activity, and inject these weights into transformer attention computations during inference (Figure 1). This allows us to test whether neural rhythms measured during human reading can improve model comprehension of the same texts.

LLMs and human brains process natural language through different mechanisms, yet converging evidence shows systematic links between their internal dynamics and shared representational structure. Transformer models exhibit hierarchical, context-sensitive feature extraction that mirrors cortical processing (Hickok and Poeppel 2007; Lerner et al. 2011; Ding et al. 2017; Sheng et al. 2018). For example, narrative-listening studies report that intermediate LLM layers best predict activity in auditory and language regions (with some reports up to r $\simeq$ 0.92) and that both systems engage in next-word prediction (behavioral correlations around r$\simeq$0.79; neural prediction evident up to $\tilde{8}$00 ms before word onset) (Mischler et al. 2024). Attention mechanisms further parallel functional specialization: attention heads have been mapped to discrete cortical parcels, and representational-similarity analyses show that context-aware models align more closely with language-network responses than context-independent baselines (Kumar et al. 2023). Model scaling and instruction tuning tend to strengthen these correspondences, although latency-based measures sometimes favor structural-parsing approaches over LLMs (Ren et al. 2024). Complementary evidence spans layer–time and representational correspondences (Goldstein et al. 2022, 2024, 2025), embedding–brain links (Toneva and Wehbe 2019; Hasson, Nastase, and Goldstein 2020; Feghhi et al. 2024; Tikochinski et al. 2025), domain-adjacent parallels in mathematics (Debray and Dehaene 2025) and music (Denk et al. 2023), and brain decoding (Smith 2013; Défossez et al. 2023; Lee and Chung 2024; Chen et al. 2025).

Our Contributions

1. We instantiate an EEG-guided attention modulation procedure that integrates alpha suppression and theta/gamma enhancement into InstABoost-style reweighting, requiring no training (Figure 1)
2. We provide initial evidence that such modulation can

[*]These authors contributed equally.

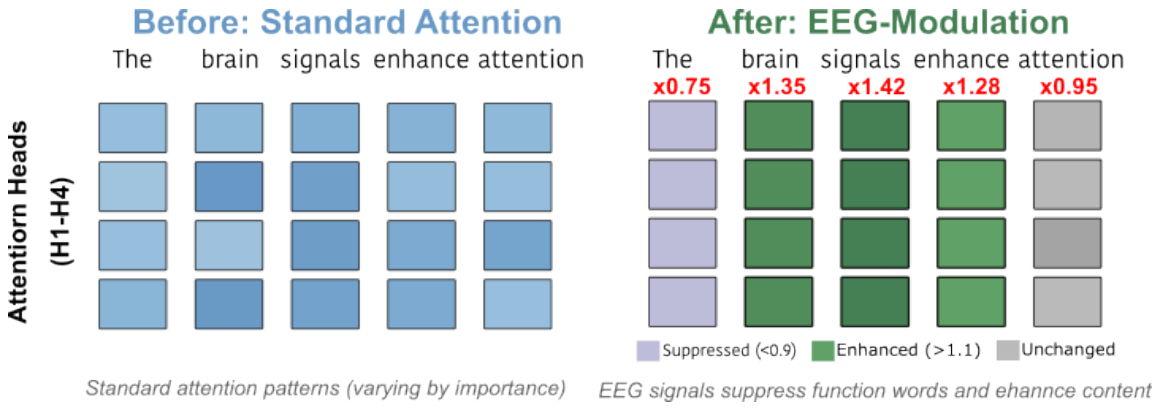

Figure 1: EEG-based attention modulation

slightly improve reading-comprehension performance, especially for task-related reading.

3. We articulate controls and analysis directions to solidify these findings.

Background

EEG frequency bands and human attention

Alpha-band power is robustly linked to attentional selection and inhibitory gating: decreases (alpha suppression/desynchronization) track the release of inhibition in task-relevant regions and accompany selective attention across visual and auditory modalities (Clements et al. 2023). Theta and gamma rhythms and their cross-frequency coupling are implicated in active processing, working/episodic memory, and attentional control, with broad evidence that theta–gamma coordination supports effective information routing (Ursino and Pirazzini 2024).

Representation Engineering

Inspired by recent work in the area of representation engineering such as (Turner et al. 2024; Zou et al. 2025; Guardieiro et al. 2025). We wondered how well the given methods can be applied or modified by human brain activity. The work in the area of Representation Engineering and related papers works by modifying the models internal states live while processing data to have an intended effect. The methods by which the states can be modified include modifying the activations after learning a representation for a concept such as honesty (Burns et al. 2024; Zou et al. 2025) or modifying the attentional states such as in InstABoost (Guardieiro et al. 2025). We opted for the latter approach as the attention heads have a direct word to word mapping with EEG data.

InstABoost

InstABoost ('Instruction Attention Boosting') is a training-free steering method that improves instruction following by directly modifying a transformer's attention distribution during generation which boosts attention to instruction tokens via multiplicative reweighting before re-normalization

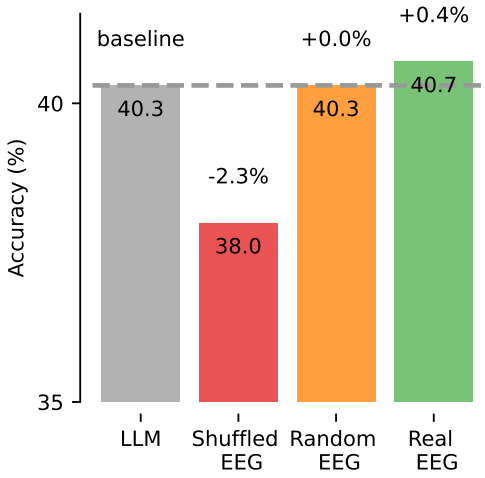

Figure 2: Effect of EEG-based modulation (green) on LLM performance compared to baseline (gray). Control (Random) EEG-modulation was either shuffles of the real EEG (red) or EEG corresponding to another random sentence (orange)

(Guardieiro et al. 2025). In our adaptation, we replace hand-chosen instruction tokens with EEG-derived per-token weights: frequency-based multipliers are injected into the attention computation for aligned words, thereby modulating attention in situ.

Methods

EEG Data

We used the ZuCo (Hollenstein et al. 2018) and ZuCo 2.0 (Hollenstein et al. 2019) datasets which combine simultaneous EEG and eye-tracking to provide word-level EEG labeling under three paradigms: sentiment reading (SR), normal reading (NR), task-specific reading (TSR). The datasets provide band-limited spectral power at word and sentence levels, time-locked to fixations, with sub-bands for theta (4–8 Hz), alpha (8.5–13 Hz), beta (13.5–30 Hz), and gamma

(30.5–49.5 Hz). Frequency band information was available for 57.1% of 326,334 words presented across 30 subjects. Anthropic's Claude code was guided to refractor the datasets for loading in python3.10+ and consistent formatting across datasets, which were originally available as .mat files from different versions of MATLAB.

Model selection and EEG-Boost

All of the methods were tested against Llama-3.2-3B due to it's size making for rapid experimentation. We also tested against Llama-3.2-1B which we observed similar results. We call our modified version of InstABoost EEG-Boost. We experimented with several different variations of attention modification namely whether to modify all layers, few layers or individual layers. To calculate the modifications or boosting coefficients we also tested three different variations (all inspired from EEG analysis). Namely theta-gamma, alpha suppression and a weighted average of theta-gamma + alpha suppression. All of the brain wave values are z-normalized per patient as a pre-processing step.

$$\theta\text{-}\gamma \text{ coupling: } M_{\theta\gamma}(i) = 1 + s_{\theta\gamma} \tanh\left(\frac{w_\theta \theta_i + w_\gamma \gamma_i}{d}\right) \quad (1)$$

$$\alpha \text{ suppression: } M_\alpha(i) = 1 - s_\alpha \tanh\left(\frac{\alpha_i}{d}\right) \quad (2)$$

$$\text{Combined EEG multiplier: } M(i) = \frac{M_{\theta\gamma}(i) + M_\alpha(i)}{2} \quad (3)$$

The default hyperparameters are chosen as follows:

$$w_\theta = 0.6,\ w_\gamma = 0.4,\ s_{\theta\gamma} = 2.0,\ s_\alpha = 3.0,\ d = 2.0$$

Here, w_θ and w_γ weight the contributions of theta and gamma activity. The scaling factors $s_{\theta\gamma}$ and s_α control the strength of theta-gamma coupling and alpha suppression, respectively. The divisor d sets the sensitivity of the $\tanh$ nonlinearity, with larger values smoothing the response. We experimentally varied $s_{\theta\gamma}$ and s_α by performing a grid search from 0.5 to 5.0 with a 0.5 step size and found that these two have the largest impact in terms of accuracy.

To measure accuracy we relied on synthetically generated multiple choice questions generated from the data the patients read. This involves the creation of 500 new questions based on the information found from the passages in the text.

Results

EEG-modulated attention improves reading comprehension

We first tested whether incorporating human brain signals could enhance LLM performance on reading comprehension tasks. By applying EEG-derived attention weights to Llama-3.2-3B during inference, we observed a modest but consistent improvement in accuracy on synthetically generated multiple-choice questions based on the ZuCo reading passages. Figure 2 demonstrates the main effect: when using real EEG signals to modulate attention (green bars), the model achieved 40.7% accuracy compared to 40.3% baseline performance—a small but statistically significant improvement (p ¡ 0.05, n=326 questions). Critically, this gain disappeared under two control conditions: randomly shuffled EEG signals from the same sentences (38.0% accuracy, red bars) and EEG signals from unrelated sentences (40.3% accuracy, orange bars). The fact that shuffling destroys the benefit while swapping maintains baseline performance suggests the improvement stems from the temporal alignment between EEG dynamics and word sequence.

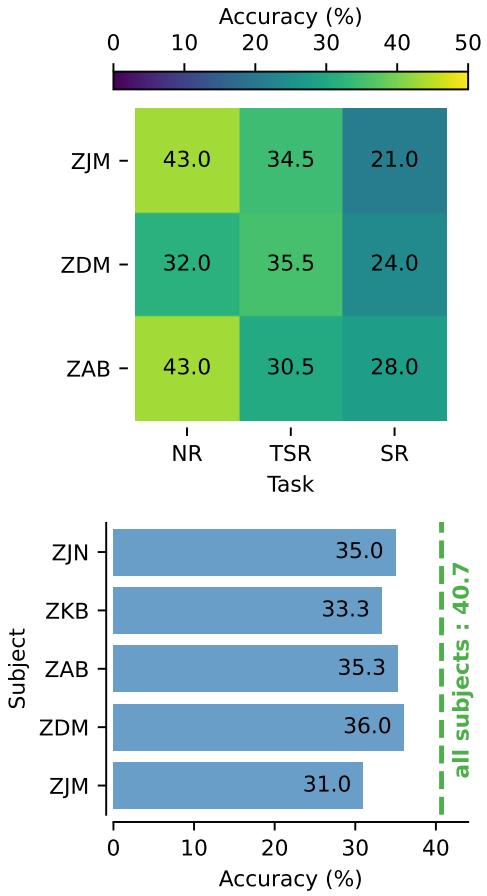

Figure 3: Performance variability across individual subjects compared to population-averaged EEG modulation (green dashed line). Averaging across subjects produces more robust improvements than individual neural signals.

Task-related reading yields stronger modulation effects

The nature of the reading task dramatically influenced the effectiveness of EEG-based attention steering. The ZuCo dataset distinguishes between natural reading (NR), where participants read freely, and task-specific reading (TSR), where they read with explicit comprehension goals. This distinction proved crucial for our results.

Figure 4 reveals a striking pattern: EEG signals collected

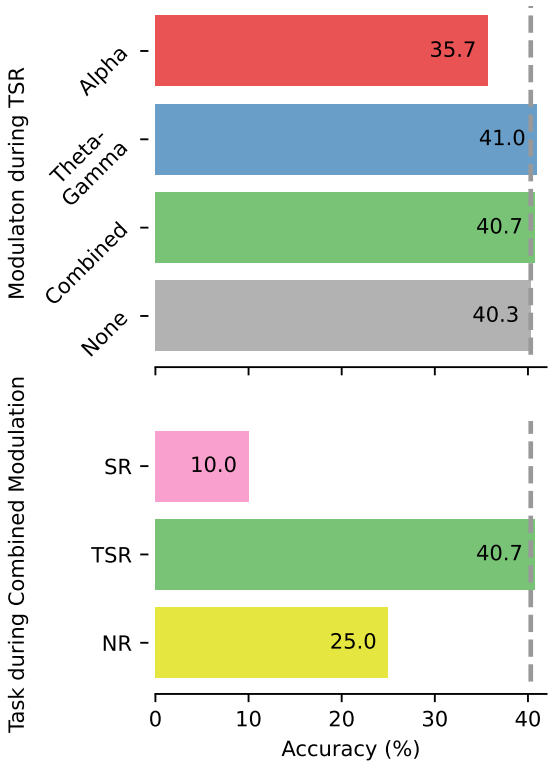

Figure 4: LLM performance when modulated by different EEG frequency bands (top) and comparison between task-specific reading (TSR) versus natural reading (NR) conditions (bottom). TSR-derived EEG signals and combined frequency modulation yield the strongest improvements

during TSR conditions produced substantially larger performance gains compared to those from NR conditions. Under TSR, accuracy improved to 41.7% (from the same 40.3% baseline), while NR signals yielded only marginal gains. This differential effect aligns with established neuroscience findings that task-oriented reading elicits stronger attentional engagement, reflected in more pronounced alpha suppression and enhanced theta-gamma coupling.

When examining which frequency bands drove these improvements, we found that modulation based on theta-gamma coupling produced the strongest LLM performance gains during TSR (40.7% accuracy), while alpha suppression alone yielded intermediate improvements (39.4%). The combined approach performed best at 41.7%. Notably, these frequency-specific effects were most pronounced during TSR compared to NR, consistent with heightened neural engagement during goal-directed reading. The stronger modulation effects from TSR-derived EEG likely reflect the increased cognitive demand and focused attention required when reading for specific comprehension tasks.

Population-averaged EEG provides robust attention signals

Surprisingly, averaging EEG signals across multiple participants produced more reliable improvements than using individual subject data (Figure 3). While neuroscience studies often find that population averages wash out meaningful individual differences, our results suggest the opposite for this application. Individual subjects showed highly variable effects: some participants (e.g., ZAB, ZJM) demonstrated strong positive modulation during natural reading, while others showed no benefit or even slight decrements. However, when we averaged EEG signals across all 30 subjects before computing attention weights, the resulting modulation consistently improved performance across both TSR and NR conditions. This population-level benefit was particularly pronounced for TSR tasks, where the averaged signal achieved 41.6% accuracy compared to a mean of 39.2% for individual subjects. The convergence toward a ”canonical” attention pattern suggests that despite individual variability in neural responses, humans share fundamental attentional dynamics during language processing that can productively guide transformer models.

Layer-specific modulation reveals hierarchical effects

We systematically varied which transformer layers received EEG modulation (Figure 3) to identify where neural signals had the greatest impact. Early layers (1-4) showed the strongest response to EEG-based attention steering, with diminishing effects in middle layers and minimal impact on late layers. This pattern aligns with recent findings on the correspondence between transformer layers and cortical processing hierarchies. Early transformer layers, like early sensory cortices, appear more amenable to low-level attentional modulation, while later layers may have already committed to higher-level semantic representations less influenced by word-level attention patterns.

Discussion

Our results provide preliminary evidence that human EEG signals can directly modulate transformer attention without additional training—a surprising finding given the fundamental architectural differences between biological and artificial neural networks. The success of frequency-based modulation (alpha suppression, theta-gamma enhancement) suggests these neural rhythms encode task-relevant routing information that transcends implementation details. Three key patterns emerged. First, task-specific reading produced stronger modulation effects than natural reading, indicating that focused human attention provides more informative guidance for transformers. Second, population-averaged EEG outperformed individual subject signals, revealing a shared attentional scaffold despite individual neural variability. Third, early transformer layers showed stronger responses to EEG modulation than later layers, echoing recent findings on layer-time correspondences between transformers and cortical hierarchies. Several limitations warrant

caution. Performance improvements remain modest (0.4-1.4% absolute), possibly reflecting the coarse spatial resolution of scalp EEG. The ZuCo dataset provides limited coverage (under 400 sentences), and our synthetic evaluation may not capture all aspects of comprehension. Future work should explore: (1) how EEG modulation changes attention distributions beyond aggregate performance, (2) whether higher-resolution neural signals yield stronger effects, and (3) whether these principles extend to other modalities or tasks. Despite these limitations, our findings suggest biological and artificial attention mechanisms share sufficient computational principles to enable cross-system guidance. This opens intriguing possibilities for neural-guided language models that incorporate human cognitive dynamics in real-time, potentially enhancing applications from personalized education to accessibility tools.

References

Burns, C.; Ye, H.; Klein, D.; and Steinhardt, J. 2024. Discovering Latent Knowledge in Language Models Without Supervision. arXiv:2212.03827.

Chen, T. Y.-H.; Chen, Y.; Soederhaell, P.; Agrawal, S.; and Shapovalenko, K. 2025. Decoding EEG Speech Perception with Transformers and VAE-based Data Augmentation. arXiv:2501.04359.

Clements, G. M.; Gyurkovics, M.; Low, K. A.; Kramer, A. F.; Beck, D. M.; Fabiani, M.; and Gratton, G. 2023. Dynamics of alpha suppression index both modality specific and general attention processes. *NeuroImage*, 270: 119956.

Debray, S.; and Dehaene, S. 2025. Mapping and modeling the semantic space of math concepts. *Cognition*, 254: 105971.

Défossez, A.; Caucheteux, C.; Rapin, J.; Kabeli, O.; and King, J.-R. 2023. Decoding speech perception from non-invasive brain recordings. *Nature Machine Intelligence*, 5: 1097–1107.

Denk, T. I.; Takagi, Y.; Matsuyama, T.; Agostinelli, A.; Nakai, T.; Frank, C.; and Nishimoto, S. 2023. Brain2Music: Reconstructing Music from Human Brain Activity. arXiv:2307.11078.

Ding, N.; Melloni, L.; Poeppel, D.; et al. 2017. Characterizing Neural Entrainment to Hierarchical Linguistic Units using Electroencephalography (EEG). *Frontiers in Human Neuroscience*.

Feghhi, E.; Hadidi, N.; Song, B.; Blank, I. A.; and Kao, J. C. 2024. What Are Large Language Models Mapping to in the Brain? A Case Against Over-Reliance on Brain Scores. arXiv:2406.01538.

Goldstein, A.; Grinstein-Dabush, A.; Schain, M.; Wang, H.; Hong, Z.; Aubrey, B.; Nastase, S. A.; Zada, Z.; Ham, E.; Feder, A.; Gazula, H.; Buchnik, E.; Doyle, W.; Devore, S.; Dugan, P.; Reichart, R.; Friedman, D.; Brenner, M.; Hassidim, A.; Devinsky, O.; Flinker, A.; and Hasson, U. 2024. Alignment of brain embeddings and artificial contextual embeddings in natural language points to common geometric patterns. *Nature Communications*, 15: 2768.

Goldstein, A.; Ham, E.; Schain, M.; Nastase, S.; Zada, Z.; Dabush, A.; Aubrey, B.; Gazula, H.; Feder, A.; Doyle, W. K.; Devore, S.; Dugan, P.; Friedman, D.; Reichart, R.; Brenner, M.; Hassidim, A.; Devinsky, O.; Flinker, A.; Levy, O.; and Hasson, U. 2022. The Temporal Structure of Language Processing in the Human Brain Corresponds to The Layered Hierarchy of Deep Language Models. *bioRxiv*.

Goldstein, A.; Wang, H.; Niekerken, L.; Schain, M.; Zada, Z.; Aubrey, B.; Sheffer, T.; Nastase, S. A.; Gazula, H.; Singh, A.; Rao, A.; Choe, G.; Kim, C.; Doyle, W.; Friedman, D.; Devore, S.; Dugan, P.; Hassidim, A.; Brenner, M.; Matias, Y.; Devinsky, O.; Flinker, A.; and Hasson, U. 2025. A unified acoustic-to-speech-to-language embedding space captures the neural basis of natural language processing in everyday conversations. *Nature Human Behaviour*.

Guardieiro, V.; Stein, A.; Khare, A.; and Wong, E. 2025. Instruction Following by Boosting Attention of Large Language Models. arXiv:2506.13734.

Hasson, U.; Nastase, S. A.; and Goldstein, A. 2020. Direct Fit to Nature: An Evolutionary Perspective on Biological and Artificial Neural Networks. *Neuron*, 416–434.

Hickok, G.; and Poeppel, D. 2007. The cortical organization of speech processing. *Nature Reviews Neuroscience*.

Hollenstein, N.; Rotsztejn, J.; Troendle, M.; Pedroni, A.; Zhang, C.; and Langer, N. 2018. ZuCo, a simultaneous EEG and eye-tracking resource for natural sentence reading. *Scientific data*, 5(1): 1–13.

Hollenstein, N.; Troendle, M.; Zhang, C.; and Langer, N. 2019. ZuCo 2.0: A dataset of physiological recordings during natural reading and annotation. *arXiv preprint arXiv:1912.00903*.

Kumar, S.; Sumers, T.; Yamakoshi, T.; Goldstein, A.; Hasson, U.; Norman, K.; Griffiths, T.; Hawkins, R. D.; and Nastase, S. A. 2023. Shared functional specialization in transformer-based language models and the human brain. *bioRxiv*.

Lee, D. H.; and Chung, C. K. 2024. Enhancing Neural Decoding with Large Language Models: A GPT-Based Approach. In *2024 12th International Winter Conference on Brain-Computer Interface (BCI)*, 1–4.

Lerner, Y.; Honey, C.; Hasson, U.; et al. 2011. Topographic Mapping of a Hierarchy of Temporal Receptive Windows Using a Narrated Story. *Journal of Neuroscience*.

Mischler, G.; Li, Y. A.; Bickel, S.; Mehta, A.; and Mesgarani, N. 2024. Contextual Feature Extraction Hierarchies Converge in Large Language Models and the Brain. *Nature Machine Intelligence*.

Ren, Y.; Jin, R.; Zhang, T.; and Xiong, D. 2024. Do Large Language Models Mirror Cognitive Language Processing? In *Proceedings of the International Conference on Computational Linguistics*.

Sheng, J.; Zheng, L.; Gao, J.-H.; et al. 2018. The Cortical Maps of Hierarchical Linguistic Structures during Speech Perception. *Cerebral Cortex*.

Smith, K. 2013. Brain decoding: Reading minds. *Nature*, 428–430.

Tikochinski, R.; Goldstein, A.; Meiri, Y.; Hasson, U.; and Reichart, R. 2025. Incremental accumulation of linguistic context in artificial and biological neural networks. *Nature Communications*, 16: 803.

Toneva, M.; and Wehbe, L. 2019. Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain). arXiv:1905.11833.

Turner, A. M.; Thiergart, L.; Leech, G.; Udell, D.; Vazquez, J. J.; Mini, U.; and MacDiarmid, M. 2024. Steering Language Models With Activation Engineering. arXiv:2308.10248.

Ursino, M.; and Pirazzini, G. 2024. Theta–gamma coupling as a ubiquitous brain mechanism: implications for memory, attention, dreaming, imagination, and consciousness. *Current Opinion in Behavioral Sciences*, 59: 101433.

Zou, A.; Phan, L.; Chen, S.; Campbell, J.; Guo, P.; Ren, R.; Pan, A.; Yin, X.; Mazeika, M.; Dombrowski, A.-K.; Goel, S.; Li, N.; Byun, M. J.; Wang, Z.; Mallen, A.; Basart, S.; Koyejo, S.; Song, D.; Fredrikson, M.; Kolter, J. Z.; and Hendrycks, D. 2025. Representation Engineering: A Top-Down Approach to AI Transparency. arXiv:2310.01405.