Is Instruction-Tuning More Brain-Aligned? Mostly a Chat-Template Artifact

Ali Asaria
Transformer Lab

Tony Salomone
Transformer Lab

Deep Gandhi
Transformer Lab
deep@lab.cloud

Abstract

A growing literature maps the internal representations of language models onto human brain responses to language, and recent work reports that *instruction-tuned* models are more "brain-aligned" than their base counterparts. We ask whether this reflects the post-training *weights* or an uncontrolled confound: the chat template that instruction-tuned models are typically fed. Using standard linear encoding models of Pereira2018 language-network fMRI, we make three contributions. First, we disentangle the two factors, comparing base and post-trained checkpoints from two families (Qwen2.5-7B; the OLMo-2-7B base→SFT→DPO→Instruct ladder) under *identical raw text*, and independently measuring the effect of applying each model's chat template. Second, we find that the post-training weight change leaves brain alignment essentially unchanged—a null across the ladder (Qwen base→Instruct p=0.92, n=10)—whereas merely applying the chat template significantly and substantially increases apparent alignment, and does so even for the *base* model, identifying the effect as an input-format artifact rather than a property of alignment training. Third, we argue this reconciles the conflicting "toward-brain" and "null" results in the recent literature, which differ largely in whether post-trained models are fed their templated format, and we draw the methodological implication that brain-alignment comparisons across models must hold input formatting fixed. Findings are from a single dataset with raw (non–noise-ceiling-normalized) correlations and n=10 subjects; multi-subject replication is the key next step.

1 Introduction

The hidden states of language models are partially predictive of human brain responses to the same language, and a linear "encoding model" fit from model features to neural signal recovers this correspondence well above chance [1, 2]. Most of this work studied *base* (pretrained) models; a natural and consequential question is whether *post-training*—instruction tuning, preference optimization (DPO), RLHF—moves a model's representations toward or away from language cortex. The recent literature disagrees: some report that instruction tuning increases brain alignment [3], while others find base and post-trained models comparably predictive [4].

We identify an uncontrolled variable that can produce an *apparent* alignment shift without any change in the weights: the *chat template*. Instruction-tuned models are trained and typically evaluated on chat-formatted inputs; feeding a post-trained model its template while feeding the base model raw text confounds the weight change with an input-distribution change. We hold input format fixed and vary one axis at a time.

Our contributions are:

1. Under *identical raw text*, the base→post-trained weight change does *not* reliably move brain-encoding alignment in either of two model families (n=10, paired Wilcoxon; §5).
2. Applying the chat template to the *same* checkpoint raises apparent alignment significantly (+0.013 to +0.019, p<0.05), ∼5–30× the weight effect, and the boost appears even in the base model—so it is an input-format effect, not an alignment-training effect (§5).
3. We show this reconciles the "toward-brain" vs. "null" literature: the sign of the reported effect tracks whether the post-trained model was fed its template (§6).

2 Related Work

Brain encoding of language models. Linear encoding models predicting fMRI/MEG from model hidden states are the field standard [1, 2, 5], with middle-to-upper layers typically most predictive and effects reported relative to a noise ceiling. Kornblith et al. [6] introduce CKA as a representation-similarity index whose properties we note when discussing metric choice, and Feather et al. [7] caution that brain-model scores must be read against an inter-subject ceiling. Methodological studies map how scale, compression, and abstraction shape alignment [8, 9, 10].

Training and alignment effects. AlKhamissi et al. [11] show untrained, randomly initialized networks already achieve substantial alignment, so the trained-vs-untrained gap is model-dependent; we adopt their untrained-control design. Several works relate the *training objective* to alignment in vision or speech [12, 5] and probe instruction-tuned multimodal models against the brain [3, 13]. Merlin and Toneva [14] and Xiao et al. [4] manipulate or compare model–brain alignment directly; the latter reports base and instruction-tuned models with comparable predictivity. Our work isolates the post-training *weight* effect from the *input-format* effect, which prior comparisons did not control.

3 Method

For a fixed set of language stimuli we extract a model's hidden states layer by layer and mean-pool over tokens to one vector per stimulus per layer. For each subject we fit a ridge-regression encoding model (RidgeCV, $\alpha \in \text{logspace}(-1, 8, 10)$) from these features to the measured per-voxel response, scoring held-out voxel-wise Pearson r under passage-grouped 5-fold cross-validation (sentences from one passage never split across train/test). We summarize a checkpoint by its mid-band-layer alignment (mean r over the central third of layers, depth-normalized so architectures with different depths are comparable). The contrast of interest is the paired difference $\Delta = r_{\text{post}} - r_{\text{base}}$, evaluated per subject and tested with a paired Wilcoxon signed-rank test across subjects. The chat-template axis applies each model's `apply_chat_template` to every stimulus and re-measures $\Delta_{\text{tmpl}} = r_{\text{templated}} - r_{\text{raw}}$ on the same checkpoint. All comparisons feed base and post-trained models *identical raw text* unless the template axis is being tested.

4 Experimental Setup

Data. We use the Pereira2018 language-network fMRI assembly (accessed via Brain-Score): 627 sentences (two experiments of 243 and 384 sentences across 96 passages), 10 subjects, with ∼1.3k

language-network voxels per subject after dropping NaN/unseen-stimulus entries. Stimuli are fed verbatim to each model.

Models. Qwen2.5-7B {base, Instruct}; the OLMo-2-1124-7B ladder {base, SFT, DPO, Instruct}; and an untrained (randomly initialized) Qwen2.5-7B of the same architecture as a control. A random-Gaussian-feature encoding provides a chance floor.

Protocol. Hidden states are extracted in `bfloat16`; encoding fits and the paired Wilcoxon tests run per subject (n=10). All experiments together used approximately 4.9 GPU-hours. Reported correlations are raw held-out r (not noise-ceiling normalized); the paired Δ contrast is within-pair on identical voxels, so it does not depend on the absolute scale.

5 Results

Sanity and anchors. Encoding is well above chance and peaks in upper-middle layers (Qwen2.5-7B base peak r=0.184 at layer 19/28; OLMo peaks near layer 15/32), matching the standard profile. The random-feature floor is $r = -0.010$, and the untrained, randomly initialized model reaches r=0.094—about half of the trained models—consistent with prior reports that architecture and tokenization alone confer substantial alignment.

Post-training weights: a null. Under identical raw text, the base→Instruct weight change does not reliably move alignment (Table 1). For Qwen the paired difference is $\Delta = -0.0003$ (p=0.92, n=10). For the OLMo ladder no stage produces a significant increase (base→SFT +0.0020, p=0.16; base→Instruct +0.0013, p=0.28); the only significant stage effect is a tiny *decrease* from SFT to DPO (−0.0007, p=0.010). Figure 1 shows base and Instruct curves overlapping at every layer.

Chat template: a significant confound. Applying each model's chat template to the same checkpoint raises apparent alignment substantially and significantly: Qwen-Instruct +0.0190 (p=0.020); OLMo-SFT/DPO/Instruct +0.0125/+0.0128/+0.0131 (each p=0.006); n=10. The effect appears even in the *base* model (Qwen-base +0.0125, p=0.064), indicating an input-format mechanism rather than an alignment-training one. The template effect is roughly 5–30× the raw weight effect (Figure 2).

6 Discussion and Limitations

Holding input format fixed, post-training weights leave brain-encoding alignment essentially unchanged, while the chat template—an input-distribution change—produces a significant, larger boost that is present even in the base model. This offers a parsimonious reconciliation of the literature: studies that fed instruction-tuned models their chat template would observe a "toward-brain" shift [3], whereas studies comparing under matched inputs would observe a null [4]. The practical implication is that brain-alignment comparisons across base and post-trained models must control the input format; otherwise a format artifact masquerades as a representational property.

We also report a methodological caution from our own pipeline: a pooled-voxel estimate suggested a +0.004 OLMo-SFT alignment gain that did *not* survive per-subject testing (+0.0020, p=0.16 at n=10); we report the per-subject numbers throughout.

Table 1: Paired per-subject effects on held-out encoding r (Pereira, n=10, paired Wilcoxon). The post-training *weight* effect is null; the *chat-template* effect is significant and far larger.

Axis	Comparison	Δr	p
Weight (raw)	Qwen base→Instruct	−0.0003	0.92
Weight (raw)	OLMo base→SFT	+0.0020	0.16
Weight (raw)	OLMo base→Instruct	+0.0013	0.28
Weight (raw)	OLMo SFT→DPO	−0.0007	0.010
Template	Qwen-Instruct (tmpl−raw)	+0.0190	0.020
Template	OLMo-SFT (tmpl−raw)	+0.0125	0.006
Template	OLMo-DPO (tmpl−raw)	+0.0128	0.006
Template	OLMo-Instruct (tmpl−raw)	+0.0131	0.006
Template	Qwen-base (tmpl−raw)	+0.0125	0.064

6.1 Threats to Validity

Internal. The template manipulation changes token statistics (added control tokens), so it measures an input-distribution effect, not specifically instruction-following; we frame it as a format confound accordingly. Encoding fits are per subject with passage-grouped folds to avoid stimulus leakage. **External.** Results are from a single dataset (Pereira2018), two model families, one extraction convention (mean-pooled hidden states), and n=10 subjects—at which the Wilcoxon test floors near p=0.06 for the smaller effects; generalization to other datasets, modalities, and aggregation choices is untested. **Construct.** We report raw held-out r without noise-ceiling normalization; absolute values therefore understate ceiling-relative alignment, though the within-pair Δ contrast is unaffected. Cross-architecture layer comparisons use depth-normalized bands rather than matched layers. **Statistics.** The reported p-values are uncorrected for multiple comparisons (nine paired tests); under a conservative Bonferroni threshold ($\alpha/9 \approx 0.0056$) the OLMo template effects (p=0.006) sit at the boundary and the Qwen-Instruct template effect (p=0.020) would not survive, so we treat the template effect as a consistent, well-powered *trend* corroborated across five model variants rather than a single decisive test—while the weight-effect null (p=0.92) is unambiguous.

7 Availability

All artifacts are available from the authors on request.[1]

8 Conclusion and Future Work

Under controlled raw text, instruction tuning and preference optimization do not move LLM–brain alignment; the apparent "more brain-aligned" effect of instruction-tuned models is largely a chat-template input-format artifact. The most important next step is replication on a high-subject-count dataset (e.g. Narratives) with noise-ceiling normalization, which would both increase statistical power beyond the n=10 ceiling here and test generalization; extending the format analysis to other input perturbations would further localize the mechanism.

[1]The models (Qwen2.5, OLMo-2) and the Pereira2018 assembly (via Brain-Score) are already publicly available; the analysis code is available on request.

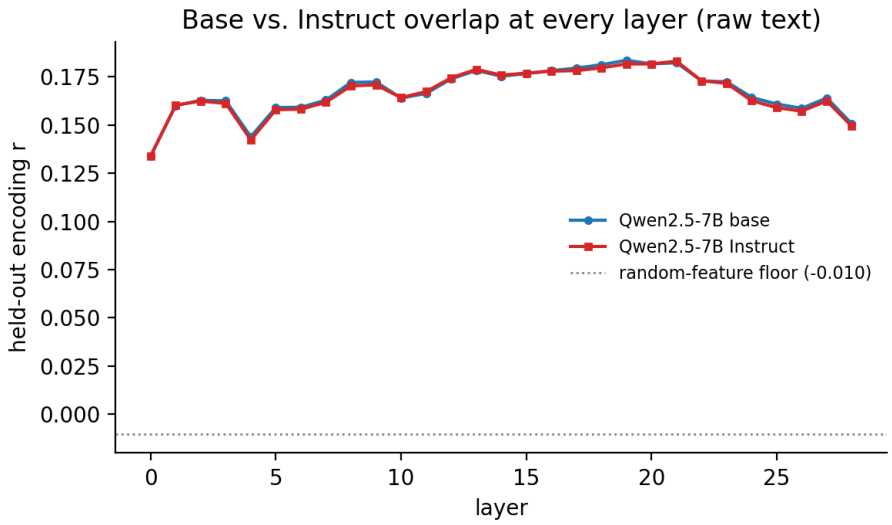

Figure 1: Per-layer held-out encoding r for Qwen2.5-7B base vs. Instruct under raw text. The curves overlap at every layer: instruction tuning does not move brain alignment when the input format is held fixed.

References

[1] Mariya Toneva and Leila Wehbe. Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain). arXiv:1905.11833 [cs.CL], 2019. URL https://arxiv.org/abs/1905.11833.

[2] Charlotte Caucheteux, Alexandre Gramfort, and Jean-Remi King. Long-range and hierarchical language predictions in brains and algorithms. arXiv:2111.14232 [q-bio.NC], 2021. URL https://arxiv.org/abs/2111.14232.

[3] Subba Reddy Oota, Khushbu Pahwa, Prachi Jindal, Satya Sai Srinath Namburi, Maneesh Singh, Tanmoy Chakraborty, Bapi S. Raju, and Manish Gupta. Task-conditioned probing of instruction-tuned multimodal LLMs: Region-specific brain alignment patterns under naturalistic stimuli. arXiv:2506.08277 [q-bio.NC], 2025. URL https://arxiv.org/abs/2506.08277.

[4] Mingqing Xiao, Kai Du, and Zhouchen Lin. Beyond representational alignment with brain-guided language models for robust reasoning. arXiv:2606.11893 [cs.LG], 2026. URL https://arxiv.org/abs/2606.11893.

[5] Juliette Millet, Charlotte Caucheteux, Pierre Orhan, Yves Boubenec, Alexandre Gramfort, Ewan Dunbar, Christophe Pallier, and Jean-Remi King. Toward a realistic model of speech processing in the brain with self-supervised learning. arXiv:2206.01685 [q-bio.NC], 2022. URL https://arxiv.org/abs/2206.01685.

[6] Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of Neural Network Representations Revisited. arXiv:1905.00414 [cs.LG], 2019. URL https://arxiv.org/abs/1905.00414.

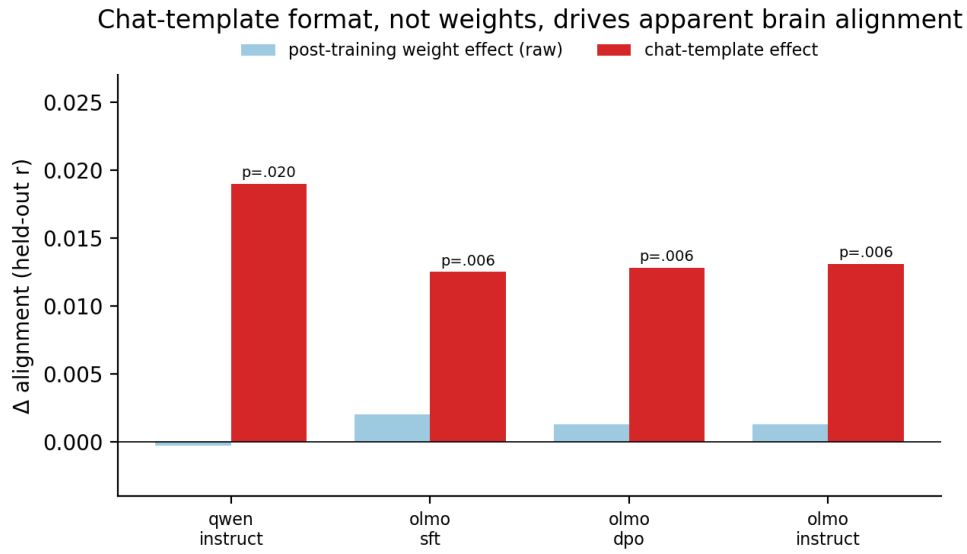

Figure 2: The post-training *weight* effect (left bars, ≈ 0, non-significant) versus the *chat-template* effect on the same checkpoints (right bars, +0.013 to +0.019, $p<0.05$). The apparent "toward-brain" shift is driven by input format, not weights.

[7] Jenelle Feather, Meenakshi Khosla, N. Apurva Ratan Murty, and Aran Nayebi. Brain-Model Evaluations Need the NeuroAI Turing Test. arXiv:2502.16238 [q-bio.NC], 2025. URL https://arxiv.org/abs/2502.16238.

[8] Subba Reddy Oota, Vijay Rowtula, Satya Sai Srinath Namburi, Khushbu Pahwa, Anant Khandelwal, Manish Gupta, Tanmoy Chakraborty, and Bapi S. Raju. Linguistic properties and model scale in brain encoding: from small to compressed language models. arXiv:2602.07547 [q-bio.NC], 2026. URL https://arxiv.org/abs/2602.07547.

[9] Emily Cheng, Aditya R. Vaidya, and Richard Antonello. Abstraction Induces the Brain Alignment of Language and Speech Models. arXiv:2602.04081 [cs.CL], 2026. URL https://arxiv.org/abs/2602.04081.

[10] Linyang He, Tianjun Zhong, Richard Antonello, Gavin Mischler, Micah Goldblum, and Nima Mesgarani. Far from the Shallow: Brain-Predictive Reasoning Embedding through Residual Disentanglement. arXiv:2510.22860 [cs.CL], 2025. URL https://arxiv.org/abs/2510.22860.

[11] Badr AlKhamissi, Greta Tuckute, Antoine Bosselut, and Martin Schrimpf. Brain-Like Language Processing via a Shallow Untrained Multihead Attention Network. arXiv:2406.15109 [cs.CL], 2024. URL https://arxiv.org/abs/2406.15109.

[12] Nils Leutenegger. Supervised Training Rapidly Degrades Early Visual Cortex Alignment Across Biologically Plausible Learning Rules. arXiv:2605.30556 [cs.LG], 2026. URL https://arxiv.org/abs/2605.30556.

[13] Subba Reddy Oota, Akshett Jindal, Ishani Mondal, Khushbu Pahwa, Satya Sai Srinath Namburi, Manish Shrivastava, Maneesh Singh, Bapi S. Raju, and Manish Gupta. Correlating instruction-

tuning (in multimodal models) with vision-language processing (in the brain). arXiv:2505.20029 [q-bio.NC], 2025. URL https://arxiv.org/abs/2505.20029.

[14] Gabriele Merlin and Mariya Toneva. When Language Models Lose Their Mind: The Consequences of Brain Misalignment. arXiv:2603.23091 [cs.CL], 2026. URL https://arxiv.org/abs/2603.23091.