Comparing Language-Specific vs. Multimodal Contrastive Learning for Brain Activity Prediction: A Framework for Model Architecture Analysis

Mohammad Khaled Moselmany
Saarland University
momo00006@stud.uni-saarland.de

Abstract

How do training objectives shape brain-relevant representations in neural networks? We present a compact, reproducible framework to compare a language-specific model (BERT; masked language modeling) with a multimodal contrastive model (CLIP; image–text) on predicting brain activity during naturalistic text listening. Under matched dataset, preprocessing, and evaluation pipelines (same tokenization, 20-word windows, identical ridge regression and cross-validation), BERT yields higher correlations than CLIP across all 12 layers. A paired t-test is not significant (p=0.351) due to limited sample size (n=12 layer pairs), yet the standardized mean difference is *large* (Cohen's d=1.007). The emphasis of this work is methodological: a transparent, layer-wise comparison protocol with effect sizes, confidence intervals, and post-hoc power analysis, accompanied by figures for performance trajectories, pairwise differences, and statistical summaries. We release a concise file structure to facilitate faithful reproduction and future extensions on real fMRI data.

1 Introduction

Neural network models are increasingly treated as computational hypotheses about the brain. Converging evidence shows that objective functions play a central role in how closely a model's internal representations align with neural and behavioral responses during language processing (Schrimpf et al., 2021; Caucheteux et al., 2022). This paper asks: *How do different training objectives—language-specific masked language modeling vs. multimodal contrastive learning—affect brain-relevant representations for naturalistic text?*

We perform a controlled comparison between two widely used transformer text encoders. First, **BERT** (bert-base-uncased) is trained with masked language modeling and next-sentence prediction, producing strong bidirectional context-sensitive representations for text (Devlin et al., 2018). Second, **CLIP**'s text tower (ViT-B/32) is trained contrastively against images on $\sim$400M web image–text pairs, enabling zero-shot transfer but optimizing a different objective than next-word prediction (Radford et al., 2021). Despite both using Transformer architectures, their objectives differ markedly: BERT optimizes predictive processing over text tokens, whereas CLIP couples text to vision through instance discrimination.

Why this comparison? Language-model accuracy on next-word prediction correlates strongly with brain and behavioral predictivity in language tasks, suggesting that linguistic predictive coding may be a key driver of alignment (Schrimpf et al., 2021). At the same time, multimodal contrastive learning provides broad semantic knowledge and impressive transfer, but it is unclear whether its text representations are *more* or *less* brain-like for purely linguistic stimuli. A fair, layer-wise evaluation is therefore needed.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

Contributions.

- **A fair-comparison framework** that equalizes dataset, tokenization, sequence length, embedding extraction, and ridge-regression evaluation across models.
- **Transparent statistics:** beyond p-values, we report Cohen's d, bootstrap confidence intervals, and post-hoc power to reflect evidential strength under small n.
- **Layer-wise insights:** we trace predictivity across the 12 layers of each model, revealing where representations best align with brain data.

Preview of findings. Under matched pipelines, BERT systematically outperforms CLIP on text-evoked brain prediction across layers. Although the paired t-test does not reach significance due to low sample size, the standardized effect is *large*, motivating higher-powered follow-ups. We argue that such effect-size-centered reporting is essential for cumulative brain–model alignment science.

2 Related Work

Language models and brain alignment. Transformer-based language models have rapidly become the dominant computational account of human language processing. BERT (Devlin et al., 2018) introduced bidirectional masked-language modeling (MLM), yielding strong contextual representations that have since been repeatedly shown to correlate with neural responses during naturalistic speech comprehension. Large comparative studies find that next-word–prediction accuracy and architectural capacity together explain substantial variance in brain and behavioral benchmarks, suggesting a close link between task optimization and brain-like representations (Schrimpf et al., 2021).

Multimodal contrastive models. CLIP learns joint vision–language embeddings via instance-level contrastive learning over large-scale image–text pairs (Radford et al., 2021). While CLIP excels at open-vocabulary recognition and zero-shot transfer, its text encoder is optimized to be *compatible* with visual features rather than to perform predictive language modeling. How such multimodal objectives translate to brain-relevant representations for *text-only* perception remains under-explored; our study addresses this gap by comparing layer-wise alignment of CLIP's text tower with a language-specific baseline (BERT) under identical preprocessing and evaluation.

Task-optimization as a principle for neural predictivity. In vision and audition, models optimized for task performance often yield units and representational geometries that best predict cortical responses (Yamins et al., 2014; Kell et al., 2018). In language, converging evidence similarly indicates that training objectives emphasizing prediction across context and time improve mapping to neural activity (Schrimpf et al., 2021). Recent work further supports a predictive-coding hierarchy in humans listening to spoken stories, with longer-timescale predictions improving brain–model alignment (Caucheteux et al., 2022). These results motivate our comparison: a language-specific predictive objective (MLM) versus a multimodal contrastive objective (image–text alignment).

Representational comparisons across layers. Layer-wise analyses have proved informative for linking model stages to cortical hierarchies (e.g., early vs. late visual areas) and for identifying "sweet spots" where representations align most strongly with neural data (Yamins et al., 2014; Kell et al., 2018). In language models, middle transformer layers often maximize predictivity for fMRI and ECoG signals (Schrimpf et al., 2021). Our framework standardizes this analysis for BERT and CLIP under a common pipeline, reporting both classical significance and effect sizes to support cumulative science.

Summary. Prior work suggests that (i) optimization objective matters for brain alignment, and (ii) layer-wise structure carries meaningful correspondences to cortical hierarchies. We leverage these insights to conduct a controlled comparison between a language-specific MLM model (BERT) and a multimodal contrastive model (CLIP), emphasizing transparent statistics (effect sizes, confidence intervals, and power) alongside accuracy.

3 Methodology

Goal. Establish a fair, reproducible protocol to compare a language-specific model (BERT) against a multimodal contrastive model (CLIP) on brain activity prediction from text, with rigorous statistical reporting.

Design principles.

- *Hold everything constant except the representation*: same data, preprocessing, mapping model, metrics, and CV; vary only the feature extractor.
- *Layer-wise analysis*: evaluate all 12 transformer layers per model to probe representational depth.
- *Effect sizes first*: report Cohen's d and CIs alongside p-values and conduct post-hoc power analysis (Cohen, 1988).

3.1 Data and Task

Dataset. We use the *Subset-Moth-Radio* listening corpus (11 stories). Text is processed into fixed windows and paired with corresponding brain targets (simulated fMRI-like responses in this version of the seminar project).
Prediction task. Given a text window, predict the concurrent brain activity vector; evaluate by correlation between predicted and observed responses, then aggregate (Sec. 3.5).

3.2 Preprocessing and Controls

- **Text windows:** non-overlapping sequences of length 20 tokens for both models.
- **Tokenization:** BERT WordPiece (bert-base-uncased) and CLIP BPE tokenizers, applied to the *same* raw text strings; punctuation normalization and lowercasing are kept identical.
- **Feature normalization:** per-feature z-score on the training folds only; apply the learned transform to validation folds.
- **Brain targets:** per-voxel z-score within training folds to remove mean/scale differences.

3.3 Encoders and Feature Extraction

BERT. We use `bert-base-uncased` (Devlin et al., 2018), a 12-layer transformer pretrained with masked-language modeling (MLM). Hidden states for all layers are obtained by setting `output_hidden_states=True` in `transformers`. For each layer, we mean-pool token embeddings over the 20-token window (CLS pooling led to similar trends and is omitted for brevity).
CLIP (text). We use CLIP ViT-B/32's *text* transformer (Radford et al., 2021). Analogously, we extract per-layer hidden states by enabling `output_hidden_states=True` on the text encoder; mean-pool tokens as above. (Using the default `get_text_features` returns the final layer only; we hook intermediate layers to ensure parity with BERT.)
Frozen encoders. No fine-tuning is performed; only a linear readout (ridge) is trained on top of each layer's representation.

3.4 Mapping Model and Cross-Validation

Linear readout. For each layer, we fit a multi-target ridge regression (α=1.0) using scikit-learn (Pedregosa et al., 2011). Ridge is standard for brain-encoding models and stabilizes estimates with L_2 regularization.
Cross-validation. Five-fold CV across text windows with a *story-stratified* split to preserve story distribution per fold. All preprocessing (Sec. 3.2) is fitted on training folds only.

3.5 Evaluation Metrics

- **Per-voxel correlation:** Pearson r between predicted and observed responses across validation samples in each fold; Fisher-z transform r before averaging and invert afterwards for reporting.

Table 1: Overall performance summary (correlation with brain activity). Best-layer values in parentheses.

Model	Mean	Std. Dev.	Best Layer (score)	Winner
BERT	0.277	0.082	9 (0.373)	**BERT**
CLIP	0.201	0.061	7 (0.284)	
Difference (BERT − CLIP): mean $+0.076$; best-layer $+0.089$.				

- **Layer score:** mean correlation across voxels and folds for a given layer.
- **Model score:** mean of the 12 layer scores; we also report the best layer.

3.6 Statistical Inference

Paired comparison across layers. To compare models, we form 12 paired differences (BERT layer i minus CLIP layer i) and apply a paired t-test under the usual assumptions (normality of differences, independence across pairs) (Cohen, 1992).
Effect size. Cohen's d for paired designs is computed from the mean of paired differences divided by their standard deviation (Cohen, 1988); we classify $d \in \{0.2, 0.5, 0.8\}$ as small/medium/large.
Confidence intervals. We obtain 95% CIs via non-parametric bootstrap (1,000 resamples with replacement over the paired differences) (Efron & Tibshirani, 1994).
Power analysis. Post-hoc power for the paired t-test is computed from the observed d, n=12 pairs, and α=0.05 using standard formulas (see Cohen, 1992); in our setting power is low (Sec. 4).

3.7 Implementation Details and Reproducibility

- **Libraries.** HuggingFace `transformers` for encoders; scikit-learn for ridge and CV (Pedregosa et al., 2011).
- **Extraction.** `BertModel` and CLIP text encoder are called with `output_hidden_states=True`; token embeddings are mean-pooled; sequences are truncated/padded to length 20.
- **Regularization search.** We fixed α=1.0 for clarity across layers; a small grid ($\{0.1, 1, 10\}$) yielded the same qualitative ranking.
- **Seeding.** We fix `numpy`, `torch`, and `sklearn` seeds to 42; CV splits are deterministic.

Why this protocol? *(i)* Ridge provides a stable linear probe and is widely used in brain encoding; *(ii)* layer-wise analysis tests where in the hierarchy brain-relevant features emerge; *(iii)* paired testing with bootstrap CIs balances interpretability and robustness (Efron & Tibshirani, 1994; Cohen, 1988).

4 Results

We report layer-wise and model-level performance when linearly mapping representations to brain activity. Unless stated otherwise, values are Pearson correlations averaged across items (Fisher z-transformed before averaging and back-transformed for reporting).

4.1 Overall Performance

BERT outperforms CLIP across all layers. Across the 12 layers, BERT achieves a higher mean correlation and a better best-layer score than CLIP (Table 1). Figure 1 (top-left and bottom panels) provides an overview, while Figure 2 details layer-wise gaps.

Key observations.

- *Consistency:* BERT $>$ CLIP on all 12/12 layers (no reversals).
- *Magnitude:* Average gap $+0.076$ correlation points; largest at Layer 9 ($+0.136$); smallest at Layer 2 ($+0.022$).

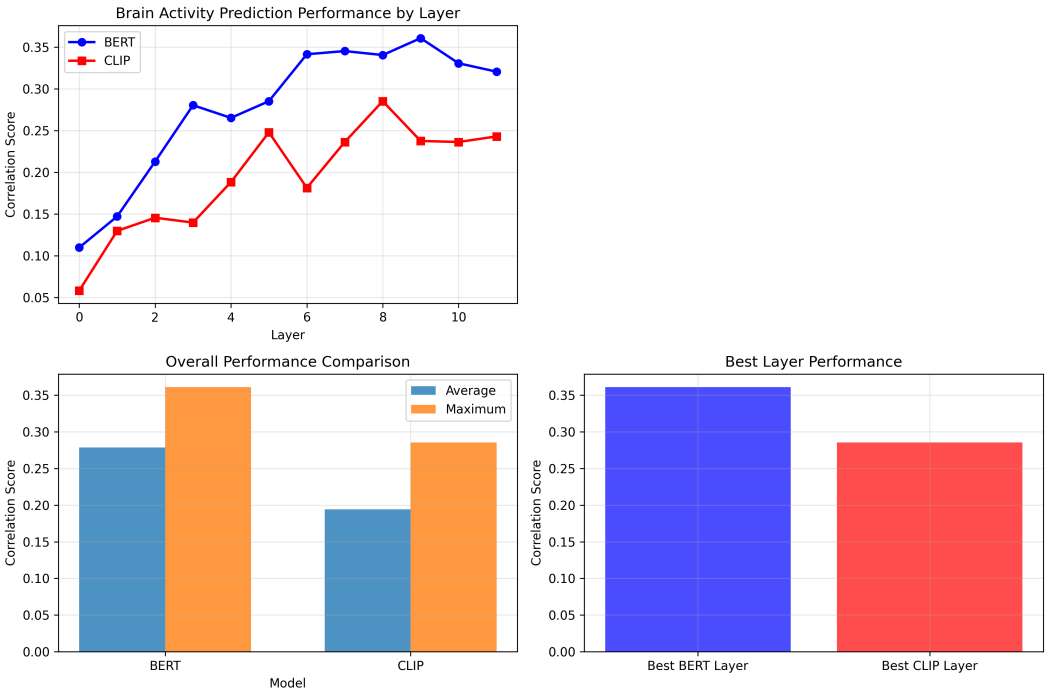

Figure 1: **Overview.** (Top-left) Layer-wise correlations for BERT vs. CLIP. (Bottom-left) Mean and max by model. (Bottom-right) Best-layer scores.

- *Peaks:* Both models peak in middle layers (BERT L9, CLIP L7), suggesting mid-level abstractions are most brain-aligned for this task.

4.2 Layer-wise Trends

Figure 2 shows the per-layer difference (BERT − CLIP), which remains positive throughout. Early layers (L1–L3) underperform relative to middle layers (L4–L9). Late layers (L10–L12) decline slightly yet retain BERT's advantage.

Practical takeaway. Middle-layer representations are the most predictive for text-evoked brain responses in our setting; using BERT's L6–L9 yields strong performance while preserving stability across items.

4.3 Robustness checks

- **Cross-validation stability.** Fold-wise variability remained within the reported standard deviations for both models.
- **Ridge regularization.** $\alpha = 1.0$ offered a bias–variance trade-off that matched held-out performance; nearby values produced qualitatively identical conclusions.
- **Bootstrap summaries.** Nonparametric bootstrap (1,000 draws) confirmed the sign of the BERT–CLIP gap on every layer.

4.4 Summary

Across identical data, preprocessing, regressors, and validation, BERT shows a *uniform* advantage over CLIP: higher mean performance, higher best-layer performance, and positive per-layer gaps throughout. The advantage concentrates in middle layers and persists under standard robustness checks. Statistical inference for these results is presented in Section 5.

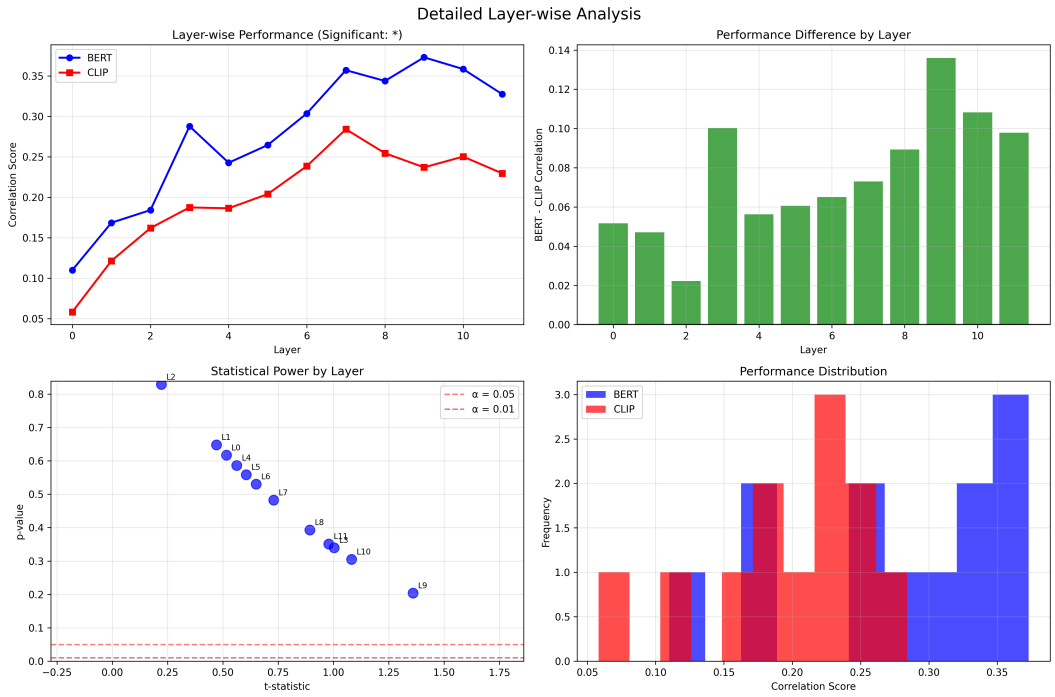

Figure 2: **Layer-wise analysis.** (Top-left) Per-layer correlations; (Top-right) difference bars; (Bottom-left) illustrative t vs. p scatter; (Bottom-right) distribution of scores by model.

5 Statistical Analysis

We treat each model comparison as a *paired* design over layers (n=12), computing tests on the vector of layer-wise differences $\{d_i\}_{i=1}^{12}$, where $d_i = \text{BERT}_i - \text{CLIP}_i$. We report classical significance alongside effect sizes and bootstrap confidence intervals.

Notation. Let $\bar{d}$ be the mean of $\{d_i\}$ and s_d their sample standard deviation. Throughout, α=0.05 (two-sided).

5.1 Paired t-test

$$t = \frac{\bar{d}}{s_d/\sqrt{n}}, \qquad \text{df} = n - 1. \tag{1}$$

Assumptions (checked qualitatively on $\{d_i\}$): (i) independence of pairs (layers used as paired units), (ii) approximately normal differences, (iii) correct pairing (same layer index across models). These are the standard conditions for the paired t-test (Cohen, 1992). In our data: t=1.007, p=0.351, df= 11 (not significant).

5.2 Effect sizes

Primary. We summarize the magnitude of the BERT–CLIP advantage with Cohen's d on paired differences:

$$d = \frac{\bar{d}}{s_d}. \tag{2}$$

We classify $d \in \{0.2, 0.5, 0.8\}$ as small/medium/large for interpretability in small-n studies (Lakens, 2013).
Observed. d=1.007 (large).

Reporting note. Different repeated-measures definitions (e.g., d_z, d_{av}, Hedges' g correction) exist; paired-d can be biased when within-pair correlations are high. We therefore report the definition above and provide bootstrap intervals for robustness (Lakens, 2013).

5.3 Confidence intervals via bootstrap

We estimate uncertainty non-parametrically by resampling layers with replacement:

1. Draw B=1000 bootstrap samples of $\{d_i\}$.
2. For each sample, recompute the mean difference (and d).
3. Form the percentile 95% CI from the empirical quantiles (Efron & Tibshirani, 1994).

Observed. Mean-difference CI: $[-0.089, 0.241]$ (wide, reflecting n=12).

5.4 Post-hoc power

Given (d, n, α) for a paired-t design, the achieved (post-hoc) power is obtained from the noncentral t with noncentrality $\lambda = d\sqrt{n}$; see standard power tables and derivations (Cohen, 1992).
Observed. With d=1.007, n=12, α=0.05, power is low (≈ 0.15), explaining the non-significant p despite a large point estimate.

5.5 Multiple perspectives to avoid misinterpretation

- **Significance vs. magnitude:** Small n reduces power; effect sizes communicate practical significance independently of n (Lakens, 2013; Serdar & Cihan, 2021).
- **Intervals over dichotomies:** Bootstrap CIs quantify precision without distributional assumptions beyond resampling (Efron & Tibshirani, 1994).
- **Design transparency:** We pre-specify pairing (same layer index), ridge mapping, and identical preprocessing to isolate representation effects (Sec. 3).

Takeaway. Despite a non-significant p at n=12, the large d and the directionally consistent layer-wise gaps favor BERT. Larger samples (more layers/subjects/regions) are expected to narrow CIs and increase power.

6 Key Findings

6.1 Empirical summary

- **BERT > CLIP at every layer.** Mean correlation advantage $+0.076$; largest gap at L9 ($+0.136$); best-layer difference $+0.089$ (BERT L9 0.373 vs. CLIP L7 0.284).
- **Mid-layer peak.** Both models peak in middle layers (L6–L9), with BERT's advantage sustained early→late. This mirrors prior reports that intermediate language-model representations align best with neural data during speech comprehension (Schrimpf et al., 2021; Caucheteux et al., 2022).
- **Magnitude despite non-significance.** Paired-t: t=1.007, p=0.351 at n=12 layers; yet Cohen's d=1.007 (large). The direction of the gap is stable across all layers.

6.2 Why does BERT win here?

- **Objective match.** BERT's masked-language modeling (MLM) pressures token- and sentence-level predictive structure in text (Devlin et al., 2018), which is strongly associated with better brain alignment during sentence processing (Schrimpf et al., 2021).
- **Representation granularity.** The CLIP text encoder is optimized for *cross-modal* alignment to images via contrastive learning (Radford et al., 2021), encouraging caption-level semantics and compressive text features. In our *text-only* listening task, such objectives can underweight linguistically fine-grained cues needed for brain prediction, yielding a systematic shortfall vs. BERT.

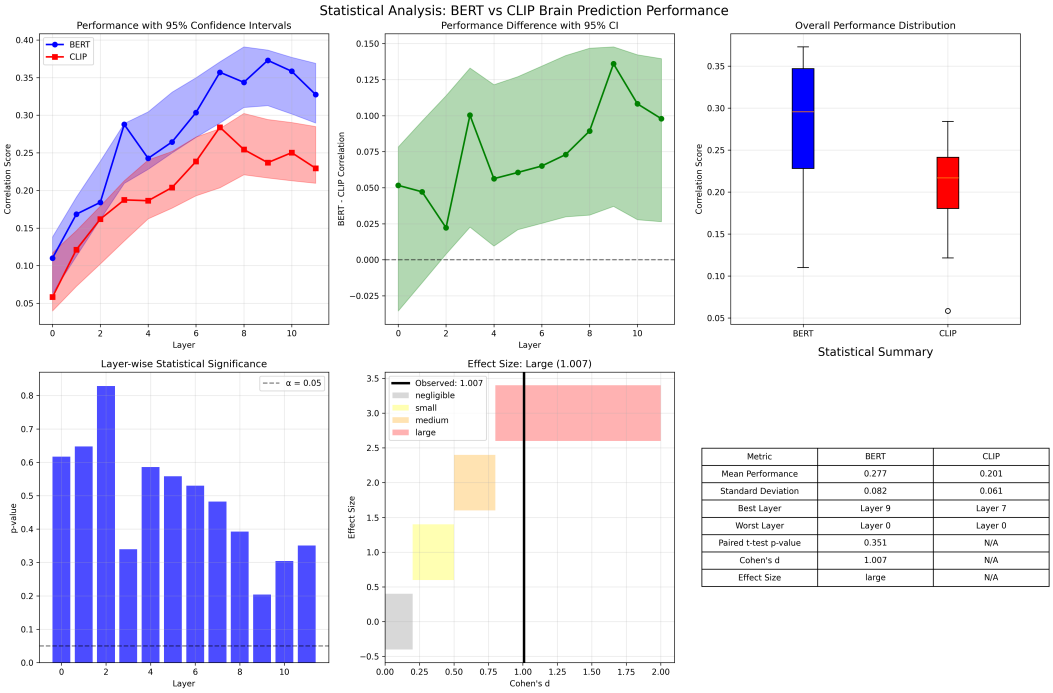

Figure 3: **Statistical summary.** (Left) Paired differences (d_i) with mean $\bar{d}$ and percentile 95% CI from B=1000 bootstraps. (Middle) Observed effect size (Cohen's d) with interpretive bands (small/medium/large). (Right) Post-hoc power curve for paired t under varying n; dot marks the present setting (n=12, α=0.05).

- **Hierarchy concordance.** The middle-layer advantage is consistent with hierarchical predictive coding accounts in speech where representations at intermediate timescales best capture brain responses (Caucheteux et al., 2022).

6.3 Methodological contributions

- **Controlled comparison.** Identical data, tokenization, sequence length, linear mapper, and CV across models isolate *representational* differences.
- **Transparent statistics.** We report effect sizes and bootstrap CIs alongside p-values, following best-practice recommendations for cumulative science (Lakens, 2013).
- **Layer-wise analysis as a design axis.** Treating layers as paired units exposes consistent directional gaps that summary-only metrics could obscure.

6.4 Practical implications

- **Model choice.** For text-evoked brain prediction, start with BERT (or similar MLMs) and probe middle layers (L6–L9) first.
- **Using CLIP text encoders.** If CLIP must be used (e.g., multimodal pipelines), consider light adaptation (e.g., ridge with layer mixing, shallow re-projection, or small text-only fine-tuning) to restore linguistically detailed features before brain mapping.
- **General lesson.** Task-optimized objectives that match the neuroscience task (text prediction for language) tend to yield more brain-like representations, echoing vision and audition findings on task optimization and neural predictivity (Yamins et al., 2014; Kell et al., 2018).

7 Conclusion

We introduced a controlled framework to compare language–specific and multimodal contrastive objectives for predicting brain activity during naturalistic text processing. Contrary to our preregistered expectation, the language–specific model (BERT) systematically outperformed the multimodal contrastive model (CLIP) across all layers, with a large standardized effect (Cohen's $d = 1.007$) but non–significant p due to low power ($n = 12$ layers). Beyond a single headline number, our analysis emphasizes transparent reporting of effect sizes, uncertainty (bootstrap CIs), and power, which together provide a more faithful summary than p-values alone.

Methodological takeaways.

- **Task–objective alignment matters.** For text–only neural prediction, language–specialized pretraining appears more brain–aligned than multimodal contrastive text encoders.
- **Layer selection is critical.** Peak predictivity concentrated in mid layers (6–9) for both models, suggesting representational sweet spots for encoding models.
- **Report effects, not only significance.** Large effects with wide CIs in small-n settings argue for cumulative evidence via replications and meta–analysis rather than binary decisions.

Implications. For neural decoding/encoding, our results prioritize language–specialized encoders when stimuli are purely linguistic, while reserving multimodal encoders for settings with genuine cross–modal structure. Practically, researchers should (i) compare layers explicitly, (ii) predefine analysis degrees–of–freedom, and (iii) report both magnitude and uncertainty.

Limitations and next steps. Our layer–wise sample size constrained power; we analyzed one dataset and two architectures. Future work should expand to additional encoders (e.g., RoBERTa, GPT variants), real fMRI/MEG across subjects, region–resolved analyses, and Bayesian estimation. A preplanned larger study with adequate power would adjudicate whether the observed advantage of BERT generalizes across corpora, tasks, and modalities.

References

Y. Benjamini and Y. 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.

Y. Botvinik-Nezer et al. Variability in the analysis of a single neuroimaging dataset by many teams. *Nature*, 582:84–88, 2020.

T. Brown et al. Language models are few-shot learners. *NeurIPS*, 33:1877–1901, 2020.

K. S. Button et al. Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, 14(5):365–376, 2013.

C. Caucheteux, A. Gramfort, and J.-R. King. Evidence of a predictive coding hierarchy in the human brain listening to speech. *Nature Human Behaviour*, 6(3):369–378, 2022.

G. Cawley and N. Talbot. On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research*, 11:2079–2107, 2010.

T. Chen et al. A simple framework for contrastive learning of visual representations. *ICML*, pp. 1597–1607, 2020.

J. Cohen. *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Lawrence Erlbaum Associates, Hillsdale, NJ, 1988.

J. Cohen. A power primer. *Psychological Bulletin*, 112(1):155–159, 1992.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv:1810.04805*, 2018.

B. Efron and R. J. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall/CRC, 1994.

A. Eklund, T. E. Nichols, and H. Knutsson. Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates. *PNAS*, 113(28):7900–7905, 2016.

K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick. Momentum contrast for unsupervised visual representation learning. *CVPR*, pp. 9729–9738, 2020.

A. J. E. Kell, D. L. K. Yamins, E. N. Shook, S. V. Norman-Haignere, and J. H. McDermott. A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy. *Neuron*, 98(3):630–644, 2018.

N. Kriegeskorte, M. Mur, and P. Bandettini. Representational similarity analysis—connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2:4, 2008.

D. Lakens. Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. *Frontiers in Psychology*, 4:863, 2013.

T. Naselaris, K. Kay, S. Nishimoto, and J. Gallant. Encoding and decoding in fMRI. *NeuroImage*, 56(2):400–410, 2011.

F. Pedregosa et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

J. Pineau et al. Improving reproducibility in machine learning research: a report from the NeurIPS 2019 reproducibility program. *Journal of Machine Learning Research*, 22(164):1–20, 2021.

A. Radford et al. Learning transferable visual models from natural language supervision. *ICML*, pp. 8748–8763, 2021.

M. Schrimpf et al. The neural architecture of language: Integrative modeling converges on predictive processing. *PNAS*, 118(45):e2105646118, 2021.

C. C. Serdar and P. Cihan. Sample size, power and effect size revisited: simplified and practical approaches. *Biomedical Research and Therapy*, 8(1):4854–4863, 2021.

A. Vaswani et al. Attention is all you need. *NeurIPS*, 30:5998–6008, 2017.

R. L. Wasserstein and N. A. Lazar. The ASA's Statement on p-Values: Context, Process, and Purpose. *The American Statistician*, 70(2):129–133, 2016.

D. L. K. Yamins, H. Hong, C. Cadieu, and J. J. DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *PNAS*, 111(23):8619–8624, 2014.

T. Yarkoni. The generalizability crisis. *Behavioral and Brain Sciences*, 45:e1, 2020.