NeuroBabyLM: Evaluating Brain Alignment of Language Models Trained on Developmentally-Plausible Datasets

Anonymous authors
Double blind review

Abstract

Abstract. Large language models (LLMs) are trained on internet-scale data, yet children acquire language from far smaller corpora. The BabyLM Challenge trains models on $\leq$100M words matching child language input, but evaluates them solely on behavioral benchmarks. We introduce **NeuroBabyLM**, which adds fMRI-based brain alignment scores to the BabyLM evaluation suite. Using ridge regression encoding models on a large-scale naturalistic listening fMRI dataset, we evaluate 15 models spanning BabyLM variants, Pythia, and LLaMA across three orders of magnitude in both size and training data. We find: (1) brain alignment follows a log-linear scaling law with model size across 14M–65B parameters; (2) T5-Base trained on 100M words outperforms Pythia-160M trained on 225B words, showing that architecture independently determines brain alignment; and (3) data-scale sensitivity is strongly architecture-dependent. These results reveal that current BabyLM models have not yet reached neural plausibility, and suggest that architectural choices may matter as much as data quantity for human-like internal representations.

Keywords: brain alignment; fMRI encoding models; BabyLM; scaling laws; developmental neuroscience

Introduction

How much language input is needed to develop brain-like language representations? Humans acquire language from roughly 10–100 million words across childhood (Warstadt et al., 2023), yet state-of-the-art LLMs require orders of magnitude more data (Kaplan et al., 2020; Hoffmann et al., 2022). The BabyLM Challenge constrains training to child-scale corpora (Warstadt et al., 2023; Hu et al., 2024), asking whether competitive language generalization can emerge from human-like data quantities. However, behavioral benchmarks alone cannot reveal whether models develop *neural plausibility*—internal representations aligned with brain activity.

Brain alignment, measured via fMRI encoding models, provides a neural evaluation axis complementary to behavioral benchmarks (Tuckute & Kanwisher, 2024; Schrimpf et al., 2021). Recent work shows alignment scales with model size in large LLMs (Antonello, Vaidya, & Huth, n.d.), and that human-like training objectives improve alignment (Tucker & Tuckute, n.d.; Aw & Toneva, 2023). Notably, Hosseini et al. (2024) found near-ceiling alignment after developmentally realistic training, while Oota et al. (2026) recently showed that neural scaling laws extend to small and compressed models—raising the question of whether BabyLM-style developmental training yields similarly aligned representations.

We present **NeuroBabyLM**: the first benchmark adding brain alignment to the BabyLM evaluation suite. Beyond benchmarking, the framework positions artificially-trained models as tools for *in-silico* developmental neuroscience—enabling controlled manipulation of training variables (architecture, data, objective) and measurement of neural consequences, circumventing the infeasibility of causal intervention in human developmental studies.

Methods

fMRI Dataset

We use the Huth lab naturalistic fMRI dataset (LeBel et al., 2023), in which participants (N=7) listened to 27 naturalistic English stories during whole-brain fMRI (TR=2 s, $\sim$80,000 voxels per subject). A subset of stories was presented 10 times to estimate the noise ceiling.

Models

We evaluate 15 models across three families (Table 1): BabyLM models (trained on 10M or 100M words) including BabyLLaMA (Timiryasov & Tastet, 2023), OPT, LTG-BERT (Samuel, Kutuzov, Velldal, & Øvrelid, 2023), and T5-Base (Raffel et al., 2020); Pythia (Biderman et al., 2023) (14M–6.9B parameters, 300B tokens); and LLaMA (7B–65B parameters, 1.4T tokens).

Encoding Model Framework

Following Han, Cho, Cha, and Lee (2025) and Antonello et al. (n.d.), we extract layer-wise hidden states for each stimulus word. Hidden states are temporally aligned to fMRI acquisition times via Lanczos interpolation, with representations from four lagged timepoints (2, 4, 6, 8 s) concatenated per TR to capture hemodynamic delay.

fMRI responses are PCA-reduced from $\sim$80,000 voxels to 512 components (Han et al., 2025), and a ridge regression encoding model maps LLM features to PCA-reduced BOLD responses on held-out stories. Performance is quantified as voxel-wise R^2, with **peak R^2 across layers** as the summary brain alignment score per model. We also report noise ceiling-normalized CC_{norm}.

Results

Scaling law. Figure 2 (left) shows a robust log-linear relationship between parameter count and peak R^2 across all models.

Table 1: Evaluated models. "w" = words; "tok" = tokens.

Family	Model	Params	Train Data
BabyLM	BabyLLaMA	58M	10M/100M w
	OPT	125M	10M/100M w
	LTG-BERT	83M	10M/100M w
	T5-Base	220M	10M/100M w
Pythia	14M–6.9B	14M–6.9B	300B tok
LLaMA	7B–65B	7B–65B	1.4T tok

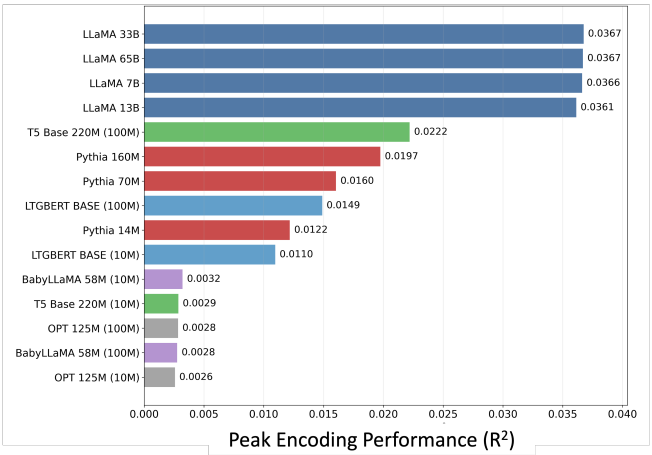

Figure 1: Peak brain alignment (R^2) for all models. BabyLM models cluster below the LLaMA ceiling. T5-Base-100M (best BabyLM, R^2=0.022) outperforms Pythia-160M (R^2=0.020) despite using ∼2000× less training data.

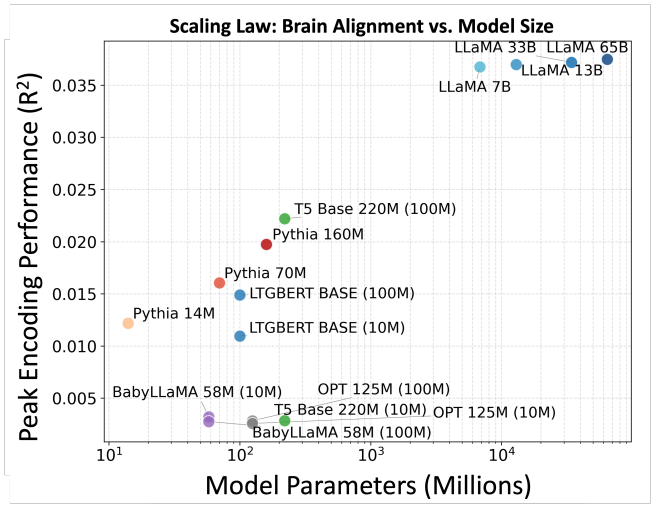

Figure 2: Brain alignment scales log-linearly with model size (14M–65B). T5-Base-100M lies above the Pythia trendline, indicating architecture-driven alignment independent of data scale.

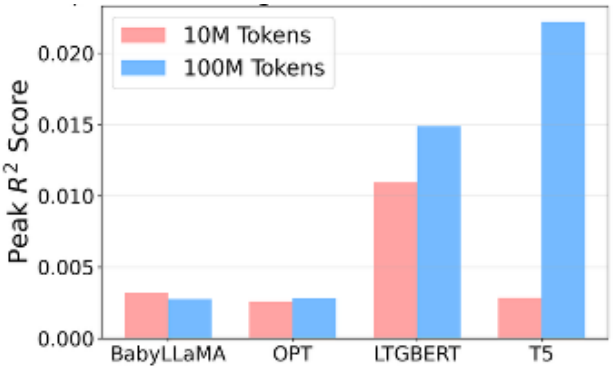

Figure 3: Effect of training data size (10M vs. 100M words) by BabyLM architecture. T5-Base gains +665%; BabyLLaMA and OPT show negligible improvement.

Within Pythia, alignment increases monotonically from 14M to 6.9B. LLaMA 33B/65B reach the ceiling (R^2=0.037). This extends the scaling laws documented in NLP (Kaplan et al., 2020; Hoffmann et al., 2022) and in brain encoding (Antonello et al., n.d.) to the developmentally-plausible regime.

Architecture matters independently of data scale. T5-Base (220M params, 100M words) achieves R^2=0.022, outperforming Pythia-160M (R^2=0.020) trained on ∼225B words (Figure 1). T5-Base also lies above the parameter-matched Pythia models in the scaling law plot (Figure 2), confirming that its encoder-decoder architecture, bidirectional attention, or span-infilling objective provides an alignment advantage independent of scale.

Data-scale sensitivity is architecture-dependent. Figure 3 shows that increasing training data from 10M to 100M words yields a +665% gain for T5-Base (R^2: 0.003→0.022), a moderate +35% for LTG-BERT, and negligible change for BabyLLaMA and OPT. This architecture-dependent sensitivity suggests only certain architectures can exploit additional training data for brain alignment.

BabyLM models have not reached neural plausibility. The best BabyLM model falls substantially below LLaMA ($\Delta R^2 \approx 0.015$), contrasting with Hosseini et al. (2024) who found near-ceiling alignment after developmentally realistic training, and with Oota et al. (2026) who found preserved alignment even in heavily compressed models. This gap may reflect the specific corpora, architectures, or training objectives used in the BabyLM Challenge.

Discussion

NeuroBabyLM reveals a tension between data efficiency and neural plausibility: models that pass BabyLM behavioral benchmarks have not yet developed brain-like representations. The strong architecture effect—T5-Base (Raffel et al., 2020) outperforming Pythia (Biderman et al., 2023) with 2000× less data—opens a tractable research question: which architectural properties (encoder-decoder structure, bidirectionality, span-infilling objective) are neurally efficient? BabyLM models are cheap to train and can be systematically varied to answer this question.

The framework enables *in-silico* developmental experiments (Chevalier-Boisvert et al., 2019): unlike human developmental neuroscience, where causal manipulation is infeasible, we can vary training curricula, architectures, and objectives while measuring neural consequences. Prior evidence that instruction tuning (Tucker & Tuckute, n.d.) and predictive objectives (Caucheteux, Gramfort, & King, 2023) improve alignment supports this approach.

Future work will expand NeuroBabyLM to all BabyLM challenge submissions, examine whether behavioral and neural

metrics dissociate across architectures, and test whether curriculum learning (Hu et al., 2024) and nonlinear multimodal encoding models (Han et al., 2025) narrow the gap between BabyLM and internet-scale LLMs.

References

Antonello, R. J., Vaidya, A. R., & Huth, A. G. (n.d.). Scaling laws for language encoding models in fMRI.

Aw, K. L., & Toneva, M. (2023, March). *Training language models to summarize narratives improves brain alignment* (No. arXiv:2212.10898). arXiv. doi: 10.48550/arXiv.2212.10898

Biderman, S., Schoelkopf, H., Anthony, Q., Bradley, H., O'Brien, K., Hallahan, E., ... van der Wal, O. (2023). Pythia: A suite for analyzing large language models across training and scaling. In *Proceedings of the 40th international conference on machine learning* (Vol. 202, pp. 2397–2430). PMLR.

Caucheteux, C., Gramfort, A., & King, J.-R. (2023, March). Evidence of a predictive coding hierarchy in the human brain listening to speech. *Nature Human Behaviour*, *7*(3), 430–441. doi: 10.1038/s41562-022-01516-2

Chevalier-Boisvert, M., Bahdanau, D., Lahlou, S., Willems, L., Saharia, C., Nguyen, T. H., & Bengio, Y. (2019, December). *BabyAI: A Platform to Study the Sample Efficiency of Grounded Language Learning* (No. arXiv:1810.08272). arXiv. doi: 10.48550/arXiv.1810.08272

Han, D. D., Cho, Y., Cha, J., & Lee, J.-Y. (2025, February). *Mind the Gap: Aligning the Brain with Language Models Requires a Nonlinear and Multimodal Approach* (No. arXiv:2502.12771). arXiv. doi: 10.48550/arXiv.2502.12771

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., ... Sifre, L. (2022). Training compute-optimal large language models. In *Advances in neural information processing systems* (Vol. 35, pp. 30016–30030). Curran Associates, Inc.

Hosseini, E. A., Schrimpf, M., Zhang, Y., Bowman, S., Zaslavsky, N., & Fedorenko, E. (2024, April). Artificial Neural Network Language Models Predict Human Brain Responses to Language Even After a Developmentally Realistic Amount of Training. *Neurobiology of Language*, *5*(1), 43–63. doi: 10.1162/nol_a_00137

Hu, M. Y., Mueller, A., Ross, C., Williams, A., Linzen, T., Zhuang, C., ... Wilcox, E. G. (2024, December). *Findings of the Second BabyLM Challenge: Sample-Efficient Pretraining on Developmentally Plausible Corpora* (No. arXiv:2412.05149). arXiv. doi: 10.48550/arXiv.2412.05149

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., ... Amodei, D. (2020). Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

LeBel, A., Wagner, L., Jain, S., Adhikari-Desai, A., Gupta, B., Morgenthal, A., ... Huth, A. G. (2023, August). A natural language fMRI dataset for voxelwise encoding models. *Scientific Data*, *10*(1), 555. doi: 10.1038/s41597-023-02437-z

Oota, S. R., Rowtula, V., Namburi, S. S. S., Pahwa, K., Khandelwal, A., Gupta, M., ... Raju, B. S. (2026). Linguistic properties and model scale in brain encoding: From small to compressed language models. *arXiv preprint arXiv:2602.07547*.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ... Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, *21*(140), 1–67.

Samuel, D., Kutuzov, A., Velldal, E., & Øvrelid, L. (2023). Trained on 100 million words and still in shape: BERT meets British National Corpus. In *Proceedings of the babylm challenge at the 27th conference on computational natural language learning* (pp. 1–14). Singapore: Association for Computational Linguistics.

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

Timiryasov, I., & Tastet, J.-L. (2023, October). *Baby Llama: Knowledge distillation from an ensemble of teachers trained on a small dataset with no performance penalty* (No. arXiv:2308.02019). arXiv. doi: 10.48550/arXiv.2308.02019

Tucker, M., & Tuckute, G. (n.d.). Increasing Brain-LLM Alignment via Information-Theoretic Compression.

Tuckute, G., & Kanwisher, N. (2024). Language in brains, minds, and machines. *Annual Review of Neuroscience*, *47*, 277–301. doi: 10.1146/annurev-neuro-120122-102046

Warstadt, A., Mueller, A., Choshen, L., Wilcox, E., Zhuang, C., Ciro, J., ... Cotterell, R. (2023). Findings of the BabyLM Challenge: Sample-Efficient Pretraining on Developmentally Plausible Corpora. In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning* (pp. 1–6). Singapore: Association for Computational Linguistics. doi: 10.18653/v1/2023.conll-babylm.1