From Language to Cognition: How LLMs Outgrow the Human Language Network

Badr AlKhamissi[1] **Greta Tuckute**[2] **Yingtian Tang**[1] **Taha Binhuraib**[3]
Antoine Bosselut[*,1] **Martin Schrimpf**[*,1]
[1]EPFL [2]MIT [3]Georgia Institute of Technology

Abstract

Large language models (LLMs) exhibit remarkable similarity to neural activity in the human language network. However, the key properties of language underlying this alignment—and how brain-like representations emerge and change across training—remain unclear. We here benchmark 34 training checkpoints spanning 300B tokens across 8 different model sizes to analyze how brain alignment relates to linguistic competence. Specifically, we find that brain alignment tracks the development of formal linguistic competence—i.e., knowledge of linguistic rules—more closely than functional linguistic competence. While functional competence, which involves world knowledge and reasoning, continues to develop throughout training, its relationship with brain alignment is weaker, suggesting that the human language network primarily encodes formal linguistic structure rather than broader cognitive functions. Notably, we find that the correlation between next-word prediction, behavioral alignment, and brain alignment fades once models surpass human language proficiency. We further show that model size is not a reliable predictor of brain alignment when controlling for the number of features. Finally, using the largest set of rigorous neural language benchmarks to date, we show that language brain alignment benchmarks remain unsaturated, highlighting opportunities for improving future models. Taken together, our findings suggest that the human language network is best modeled by formal, rather than functional, aspects of language.[1]

1 Introduction

Deciphering the brain's algorithms underlying our ability to process language and communicate is a core goal in neuroscience. Human language processing is supported by the brain's language network (LN), a set of left-lateralized fronto-temporal regions in the brain (Binder et al., 1997; Bates et al., 2003; Gorno-Tempini et al., 2004; Price, 2010; Fedorenko, 2014; Hagoort, 2019) that respond robustly and selectively to linguistic input (Fedorenko et al., 2024a). Driven by recent advances in machine learning, large language models (LLMs) trained via next-word prediction on large corpora of text are now a particularly promising model family to capture the internal processes of the LN. In particular, when these models are exposed to the same linguistic stimuli (e.g., sen-

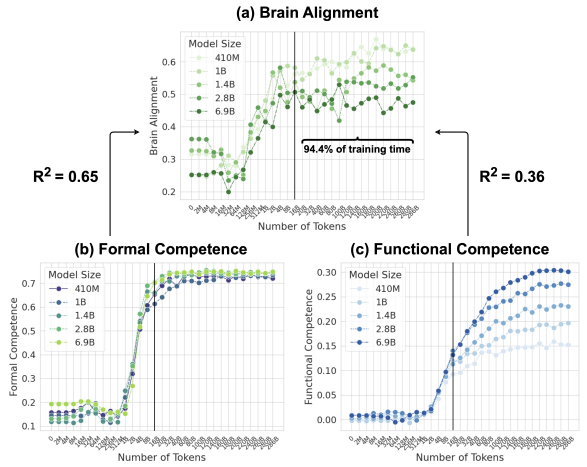

Figure 1: **Model Alignment with the Human Language Network is Primarily Driven by Formal than Functional Linguistic Competence. (a)** Average brain alignment across five Pythia models and five brain recording datasets, normalized by cross-subject consistency, throughout training. **(b)** Average normalized accuracy of the same models on formal linguistic competence benchmarks (two benchmarks). **(c)** Average normalized accuracy on functional linguistic competence benchmarks (six benchmarks). The x-axis is logarithmically spaced up to 16B tokens, capturing early training dynamics, and then evenly spaced every 20B tokens from 20B to ~300B tokens.

[*] Equal Supervision
[1]Project Page: language-to-cognition.epfl.ch

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 24333–24351
November 4-9, 2025

tences or narratives) as human participants during neuroimaging and electrophysiology experiments, they account for a substantial portion of neural response variance (Schrimpf et al., 2021; Caucheteux and King, 2022; Goldstein et al., 2022; Pasquiou et al., 2022; Aw et al., 2023; Tuckute et al., 2024a; AlKhamissi et al., 2025; Rathi et al., 2025).

1.1 Key Questions and Contributions

This work investigates four key questions, all aimed at distilling *why* LLM aligns to brain responses. Specifically, we investigate the full model development cycle as a combination of model architecture (structural priors) and how linguistic competence emerges across training (developmental experience). We ask: (1) What drives brain alignment in untrained models? (2) Is brain alignment primarily linked to formal or functional linguistic competence (Mahowald et al., 2024)? (3) Do language models diverge from humans as they surpass human-level prediction? (4) Do current LLMs fully account for the explained variance in brain alignment benchmarks? To answer these questions, we introduce a rigorous brain-scoring framework to conduct a controlled and large-scale analysis of LLM brain alignment.

Our findings reveal that the initial brain alignment of models with untrained parameters is driven by context integration. During training, alignment primarily correlates with formal linguistic competence—tasks that probe mastery of grammar, syntax, and compositional rules, such as identifying subject–verb agreement, parsing nested syntactic structures, or completing well-formed sentences. This competence saturates relatively early in training ($\sim$ 4B tokens), consistent with a plateauing of model-to-brain alignment. Functional linguistic competence, in contrast, concerns how language is used in context to convey meaning, intent, and social/pragmatic content—for example, tasks involving discourse coherence, reference resolution, inference about speaker meaning, or interpreting figurative language. Functional competence emerges later in training, tracks brain alignment less strongly, and continues to grow even after alignment with the language network has saturated.

This disconnect later in training is further exemplified by a fading of the correlation between models' brain alignment and their next-word-prediction performance, as well as their behavioral alignment. Further, we show that model size is not a reliable predictor of brain alignment when controlling for the number of features, challenging the assumption that larger models necessarily resemble the brain more. Finally, we demonstrate that current brain alignment benchmarks remain unsaturated, indicating that LLMs can still be improved to model human language processing.

2 Preliminaries & Related Work

A Primer on Language in the Human Brain The human language network (LN) is a set of left-lateralized frontal and temporal brain regions supporting language. These regions are functionally defined by contrasting responses to language inputs over perceptually matched controls (e.g., lists of non-words) (Fedorenko et al., 2010). The language network exhibits remarkable selectivity for language processing compared to various non-linguistic inputs and tasks, such as music perception (Fedorenko et al., 2012; Chen et al., 2023) or arithmetic computation (Fedorenko et al., 2011; Monti et al., 2012) (for review, see Fedorenko et al. (2024a)) and the language network only shows weak responses when participants comprehend or articulate meaningless non-words (Fedorenko et al., 2010; Hu et al., 2023). This selectivity profile is supported by extensive neuroimaging research and further corroborated by behavioral evidence from aphasia studies: when brain damage is confined to language areas, individuals lose their linguistic abilities while retaining other skills, such as mathematics (Benn et al., 2013; Varley et al., 2005), general reasoning (Varley and Siegal, 2000), and theory of mind (Siegal and Varley, 2006).

Model-to-Brain Alignment Prior work has shown that the internal representations of certain artificial neural networks resemble those in the brain. This alignment was initially observed in the domain of vision (Yamins et al., 2014; Khaligh-Razavi and Kriegeskorte, 2014; Cichy et al., 2016; Schrimpf et al., 2018, 2020; Cadena et al., 2019; Kubilius et al., 2019; Zhuang et al., 2021) and has more recently been extended to auditory processing (Kell et al., 2018; Tuckute et al., 2023; Koumura et al., 2023) and language processing (Schrimpf et al., 2021; Caucheteux and King, 2022; Goldstein et al., 2022; Kauf et al., 2023; Hosseini et al., 2024; Aw et al., 2023; AlKhamissi et al., 2025; Tuckute et al., 2024b; Rathi et al., 2025).

Untrained Models Recent work in vision neuroscience has shown that untrained convolutional

networks can yield high brain alignment to recordings in the visual ventral stream without the need for training (Geiger et al., 2022; Kazemian et al., 2024). Other works have investigated the inductive biases in different architectures and initializations in models of visual processing (Cichy et al., 2016; Cadena et al., 2019; Geiger et al., 2022), speech perception (Millet and King, 2021; Tuckute et al., 2023), and language (Schrimpf et al., 2021; Pasquiou et al., 2022; Hosseini et al., 2024), highlighting that randomly initialized networks are not random functions (Teney et al., 2024).

3 Methods

3.1 Benchmarks for Brain Alignment

Neuroimaging & Behavioral Datasets The neuroimaging datasets used in this work can be categorized along three dimensions: the imaging modality, the context length of the experimental materials, and the modality through which the language stimulus was presented to human participants (auditory or visual). Table 1 in Appendix A provides an overview of all datasets in this study. To focus specifically on language, we consider neural units (electrodes, voxels, or regions) associated with the brain's language network, as localized by the original dataset authors using the method described in the Section 3.2 and implemented in Brain-Score (Schrimpf et al., 2020, 2021) (however, see Appendix J for control brain regions). An exception is the NARRATIVES dataset, which lacks functional localization. We here approximate the language regions using a probabilistic atlas of the human language network (Lipkin et al., 2022), extracting the top-10% language-selective voxels (from the probabilistic atlas) within anatomically defined language parcels, in line with the functional localization procedure used in the other datasets. In an additional analysis, we investigate model alignment with language behavior using the (Futrell et al., 2018) dataset, which contains self-paced, per-word human reading times. See Appendix A for details of each dataset. To the best of our knowledge, this study examines the largest number of benchmarks compared to previous work, providing a more comprehensive and reliable foundation for identifying the properties that drive brain alignment in LLMs. The diversity of datasets ensures that our conclusions generalize beyond specific experimental stimuli and paradigms.

Brain-Alignment Metrics Following standard practice in measuring brain alignment, we train a ridge regression model to predict brain activity from model representations, using the same linguistic stimuli presented to human participants in neuroimaging studies (Schrimpf et al., 2020, 2021). We then measure the Pearson correlation between the predicted brain activations and the actual brain activations of human participants on a held-out set that covers entirely different stories or topics (see Section 4). This process is repeated over k cross-validation splits, and we report the average (mean) Pearson correlation as our final result. We refer to this metric as Linear Predictivity. In Section 5.1, we demonstrate why other metrics such as Centered Kernel Alignment (CKA; Kornblith et al., 2019) and Representational Similarity Analysis (RSA; Kriegeskorte et al., 2008) are not suitable measures for brain alignment on current language datasets.

Estimation of Cross-Subject Consistency To assess the reliability of our datasets and account for the inherent noise in brain recordings, we compute a cross-subject consistency score (Feather et al., 2025), also referred to as the noise ceiling (Schrimpf et al., 2021). The consistency score is estimated by predicting the brain activity of a held-out subject using data from all other subjects, through 10-fold cross-validation of all subjects. To obtain a conservative ceiling estimate, we extrapolate subject pool sizes and report the final value based on extrapolation to infinitely many subjects. For TUCKUTE2024 we use the theoretical estimate provided by (Tuckute et al., 2024b). Consistency scores are provided in Appendix K. To aggregate scores across benchmarks, we normalize each model's Pearson correlation (r) score for Linear Predictivity by the cross-subject consistency estimate, using the formula: (normalized score $= \frac{\text{raw score}}{\text{consistency}}$). The final alignment score for each model is reported as the average across all benchmarks. Otherwise, when reporting raw alignment, we compute the mean Pearson correlation across datasets without normalization.

3.2 Functional Localization

The human language network (LN) is defined *functionally* which means that units are chosen according to a 'localizer' experiment (Saxe et al., 2006). Specifically, the LN is the set of neural units (e.g., voxels/electrodes) that are more selective to sentences over a perceptually-matched control condi-

tion (Fedorenko et al., 2010). When selecting units from artificial models for comparison against LN units, previous work selected output units from an entire Transformer block based on brain alignment scores (Schrimpf et al., 2021). However, LLMs learn diverse concepts and behaviors during their considerable pretraining, not all of which are necessarily related to language processing, e.g., storage of knowledge (AlKhamissi et al., 2022) and the ability to perform complex reasoning (Huang and Chang, 2023). Therefore, we here follow the method proposed by (AlKhamissi et al., 2025) that identifies language units in LLMs using functional localization as is already standard in neuroscience. This approach offers a key advantage: it enables direct comparisons across models by selecting a fixed set of units, identified through the independent localizer experiment. In this work, we localize 128 units for all models unless otherwise specified, and we show in Appendix H that the results hold when selecting a different number of units.

3.3 Benchmarks for Linguistic Competence

There is substantial evidence in neuroscience research that formal and functional linguistic competence are governed by distinct neural mechanisms (Mahowald et al., 2024; Fedorenko et al., 2024a,b). Formal linguistic competence pertains to the knowledge of linguistic rules and patterns, while functional linguistic competence involves using language to interpret and interact with the world. Therefore, to accurately track the evolution of each type of competence during training, we focus on benchmarks that specifically target these cognitive capacities in LLMs.

Formal Linguistic Competence To assess formal linguistic competence, we use two benchmarks: BLiMP (Warstadt et al., 2019) and SyntaxGym (Gauthier et al., 2020). BLiMP evaluates key grammatical phenomena in English through 67 tasks, each containing 1,000 minimal pairs designed to test specific contrasts in syntax, morphology, and semantics. Complementing this, SyntaxGym consists of 31 tasks that systematically measure the syntactic knowledge of language models. Together, these benchmarks provide a robust framework for evaluating how well LLMs acquire and apply linguistic rules.

Functional Linguistic Competence Functional competence extends beyond linguistic rules, engaging a broader set of cognitive mechanisms. To assess this, we use six benchmarks covering world knowledge (ARC-Easy, ARC-Challenge (Clark et al., 2018)), social reasoning (Social IQa (Sap et al., 2019)), physical reasoning (PIQA (Bisk et al., 2019)), and commonsense reasoning (WinoGrande (Sakaguchi et al., 2019), HellaSwag (Zellers et al., 2019)). Together, these benchmarks provide a comprehensive evaluation of an LLM's ability to reason, infer implicit knowledge, and navigate real-world contexts.

Metrics Inline with prior work, we evaluate all benchmarks in a zero-shot setting, using surprisal as the evaluation metric. where the model's prediction is determined by selecting the most probable candidate, as packaged in the language model evaluation harness (Gao et al., 2024). We report accuracy normalized by chance performance, where 0% indicates performance at the random chance level.

Benchmark for Language Modeling We use a subset of FineWebEdu (Penedo et al., 2024) to evaluate the perplexity of the models on a held-out set. Specifically, use a maximum sequence length of 2048, and evaluate on the first 1000 documents of the `CC-MAIN-2024-10` subset.

3.4 Large Language Models (LLMs)

Throughout this work, we use eight models from the Pythia model suite (Biderman et al., 2023), spanning a range of sizes: {14M, 70M, 160M, 410M, 1B, 1.4B, 2.8B, 6.9B}. Each model is evaluated across 34 training checkpoints, spanning approximately 300B tokens. These checkpoints include the untrained model, the final trained model, and 16 intermediate checkpoints that are logarithmically spaced up to 128B tokens. The remaining 14 checkpoints are evenly spaced every 20B tokens from 20B to 280B tokens, ensuring a comprehensive analysis of alignment trends throughout training. Since smaller models fail to surpass chance performance on many functional benchmarks, we exclude 14M, 70M, 160M from analyses that compare brain alignment with functional performance.

4 Rigorous Brain-Scoring

While substantial progress has been made in measuring alignment between LLM representations and neural activity, there's no standard for comparing brain alignment across datasets and conditions. Therefore, to ensure we perform meaningful inferences, we propose two criteria: (1) alignment

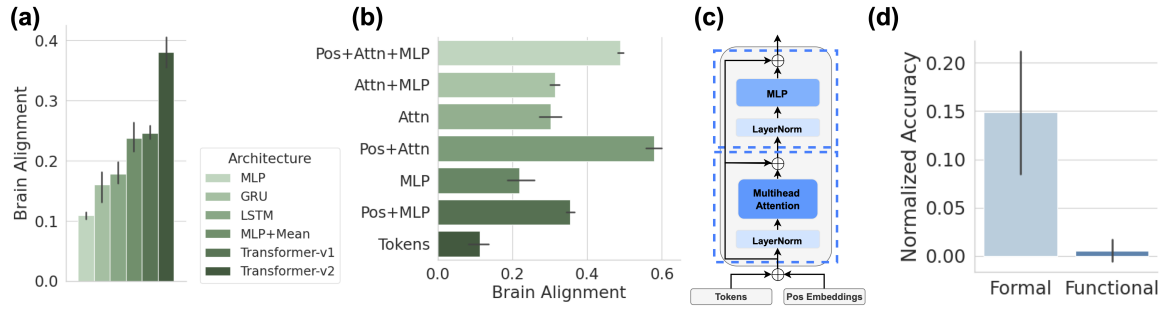

Figure 2: **Context Integration drives Brain Alignment of Untrained Models**. **(a)** Sequence-based models (GRU, LSTM, Transformers, and mean pooling) achieve higher brain alignment than models that rely solely on the last token representation (Linear, MLP), highlighting the importance of temporal integration. Error bars report five random initializations in all subplots. **(b)** Ablation study of architectural components in a single untrained TRANSFORMER-V2 block, demonstrating that attention mechanisms combined with positional encoding yield the highest brain alignment. **(c)** Diagram of the Transformer block architecture used in (b), with components grouped into attention (lower box) and MLP (upper box). **(d)** The average performance of five Pythia models with untrained parameters on formal and functional linguistic competence benchmarks, showing that formal competence exceeds chance level even in untrained parameter models.

should reflect stimulus-driven responses, dropping for random token sequences; and (2) models should generalize to new linguistic contexts. We justify our metrics and cross-validation choices accordingly. For all benchmarks, we identify language-selective units to ensure fair model comparisons, consistent with neural site selection in neuroscience (AlKhamissi et al., 2025).

4.1 Robust Metrics and Generalization Tests

Measuring Stimulus-Driven Responses We first ask if the alignment procedure is meaningful, i.e., whether the encoding models capture meaningful linguistic information and generalize to new linguistic contexts. Figure 6(a) in Appendix B shows average brain alignment across all brain datasets under three conditions: (1) a pretrained model processing original stimuli, (2) a pretrained model processing random token sequences, and (3) an untrained model processing original stimuli. To evaluate metric reliability, we expect random sequences to yield significantly lower alignment than real stimuli. However, CKA fails this criterion, assigning similar alignment scores to both, and even untrained models surpass pretrained ones. In contrast, linear predictivity differentiates between real and random stimuli, more so than RSA.

Generalization and Contextualization The second criterion we propose is that LLMs with high brain alignment should be able to generalize to held-out stimuli, with a preference for generalizing far outside the stimuli used for mapping the model to brain activity. A key factor in designing a corresponding cross-validation scheme is contextualization—how the data is split into train and test sets (Feghhi et al., 2024). The PEREIRA2018 dataset consists of 24 topics composed of multi-sentence passages, and sentences are presented in their original order to both humans and models. A random sentence split (contextualization) allows sentences from the same topic in both train and test sets, and is thus less demanding of generalization. A stronger generalization test ensures entire topics are held out, preventing models from leveraging shared context. Figure 6(b) shows that contextualization makes it easier for the model to predict brain activity. In contrast, topic-based splits halve the raw alignment score for pretrained models. The score of untrained models is reduced even more strongly when enforcing generalization across topics, suggesting that much of their alignment is context-dependent. Nonetheless, untrained models retain significant alignment – about 50% of pretrained models – even with strong generalization requirements.

5 Results

The following sections progressively unpack the emergence and limits of brain alignment with the human language network in LLMs. Section 5.1 establishes the foundation by showing that untrained models already exhibit modest brain alignment, pointing to the role of architectural priors. Building on this, Section 5.2 tracks how alignment evolves

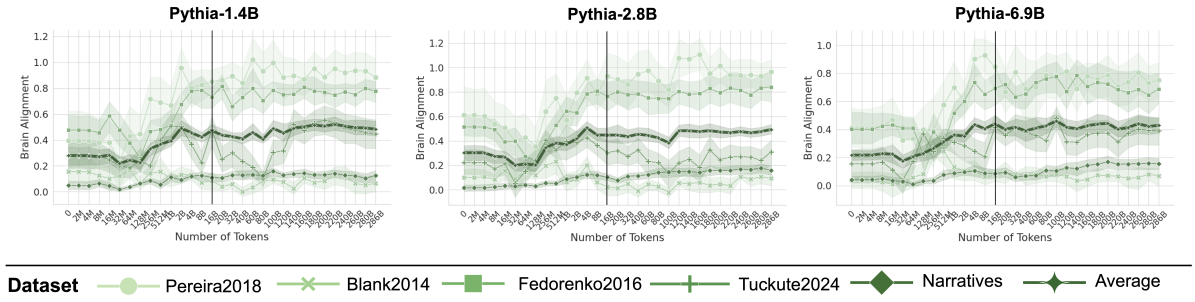

Figure 3: **Brain Alignment Saturates Early on in Training.** Plots indicate the brain alignment scores of three models from the Pythia model suite with varying sizes (log x-axis up to 16B tokens, uneven spacing after black line). Scores are normalized by their cross-subject consistency scores. Alignment quickly peaks around 2–8B tokens before saturating or declining, regardless of model size (see Appendix D and F for more models).

with training and reveals that it strongly correlates with the early acquisition of formal linguistic competence, but less so with functional abilities. Section 5.3 then shows that as models exceed human-level performance in next-word prediction, their brain and behavioral alignment begins to diverge, suggesting that at this point, LLMs outgrow their initial alignment with human language processing.

5.1 Brain Alignment of Untrained Models

In Figure 6 we show that untrained models, despite achieving lower alignment scores than their pretrained counterparts ($\sim 50\%$), still achieve relatively decent alignment and surpass that of the models evaluated with a random sequence of tokens. Therefore, we here ask, what are the main drivers for this surprising alignment.

Inductive Biases of Untrained Models We evaluate the brain alignment of various LLMs with untrained parameters to determine which architecture exhibits the strongest inductive bias toward the human language network. Figure 2(a) presents the average alignment across five different random initializations for six different untrained models. Each model consists of a stack of two building blocks from its respective architecture, with a hidden state of 1024. To ensure a fair comparison, we apply the localizer to the output representations of the last token in the sequence from these two blocks, extracting 128 units to predict brain activity. Our findings reveal two key insights. First, sequence-based models—such as GRU, LSTM, TRANSFORMERS, and even a simple mean operation over token representations—exhibit higher brain alignment than models that rely solely on the last token's representation, such as LINEAR or MLP. In other words, context or temporal integration is a crucial factor in achieving high alignment. Second, we observe a notable difference between TRANSFORMER-V1 and TRANSFORMER-V2. While TRANSFORMER-V2 applies static positional embeddings by directly adding them to token embeddings, TRANSFORMER-V1 uses rotary position encoding. Our results suggest that static positional encoding enables models to capture intrinsic temporal dynamics in sentences—possibly tracking evolving word positions—providing further evidence that temporal integration is critical for brain-like language representations.

Key Components of Transformers To further isolate the key elements responsible for brain alignment in untrained parameter models, we perform an ablation study on the architectural components of TRANSFORMER-V2 using a single block (Figure 2(c)). By focusing on the untrained model, we isolate the effect of architecture alone, without confounding influences from training. The architectural components analyzed are labeled on the left of each bar in Figure 2(b). `Attn` refers to all components inside the lower box in Figure 2(c), including the first layer norm, multi-head attention, and the residual connection that follows. `MLP` corresponds to the components in the upper box, comprising the post-attention layer norm, MLP, and the subsequent residual layer. `Pos` represents the addition of positional embeddings to token embeddings. `Tokens` means the model directly returns the raw token embeddings without further processing. This systematic ablation helps pinpoint the components that contribute most to brain alignment. Once again, we observe that integration across tokens, via attention mechanisms and positional encoding, yields the highest brain alignment. Further, we found that untrained parameter models

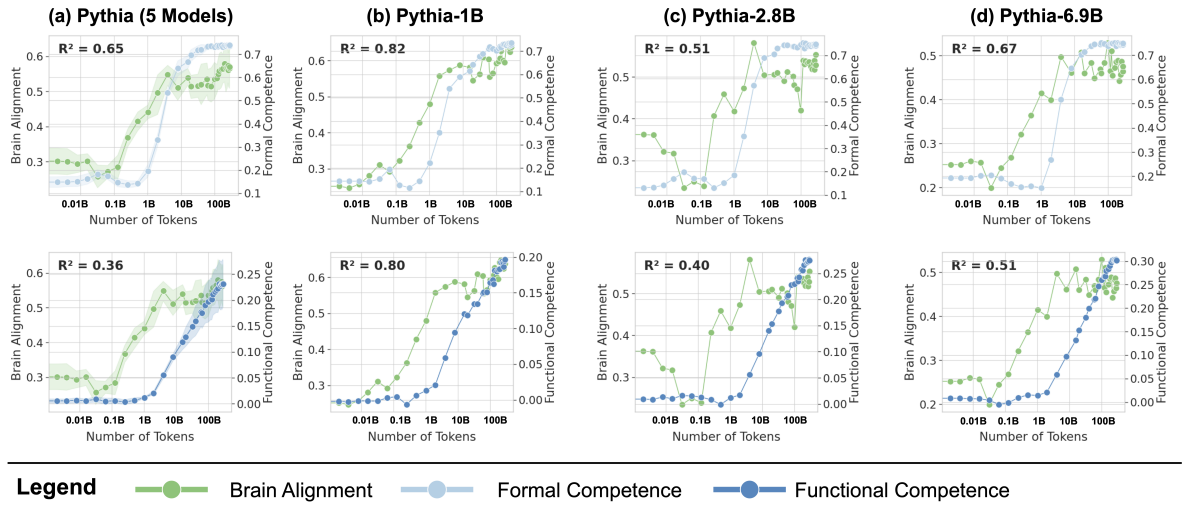

Figure 4: **Formal Competence Tracks Brain Alignment More Closely Than Functional Competence.** Each column compares how the evolution of formal competence (top) and functional competence (bottom) tracks the evolution of brain alignment during training. The R^2 values quantify the strength of this relationship, with higher values in formal competence suggesting it as the key driver of the observed brain alignment. **(a)**: The data averaged across models of five different sizes. **(b-d)**: the same comparison as in (a), but with comparisons were made for models from the Pythia suite with three different sizes.

perform better than chance-level performance on formal competence benchmarks, mirroring their non-zero brain alignment. In contrast, functional competence benchmarks remain at chance level for untrained models. This further supports the finding that brain alignment is primarily driven by formal, rather than functional, linguistic competence. (see Figure 2(d)).

5.2 Brain Alignment Over Training

Having established the architectural components that make an untrained model brain-aligned in the previous section, we now investigate how brain alignment evolves during training. To do so, we use the Pythia model suite (Biderman et al., 2023), which consists of models of various sizes, all trained on the same ∼300B tokens, with publicly available intermediate checkpoints. We report results for a model from a different family, SmolLM2-360M (Allal et al., 2025), which provides checkpoints at 250B-token intervals, in Appendix F.

Figure 3 illustrates the brain alignment of six Pythia models across five brain recording datasets at 34 training checkpoints, spanning approximately 300B tokens. Each panel presents checkpoints that are logarithmically spaced up to the vertical line, emphasizing the early-stage increase in brain alignment, which occurs within the first 5.6% of training time. Beyond this point, the panels display the remaining training period, where brain alignment stabilizes. More specifically, we observe the following trend: (1) Brain alignment is similar to the untrained model until approximately 128M tokens. (2) A sharp increase follows, peaking around 8B tokens. (3) Brain alignment then saturates for the remainder of training. Despite the vast difference in model sizes shown in Figure 3, the trajectory of brain alignment is remarkably similar.

Alignment Tracks Formal Competence Following the observation that brain alignment plateaus early in training, we next investigate how this relates to the emergence of formal and functional linguistic competence in LLMs. Figure 4 displays the average brain alignment alongside the average performance on formal competence benchmarks (top row) and functional competence benchmarks (bottom row). This is shown for three Pythia models (1B, 2.8B, and 6.9B parameters) and the average of five Pythia models (first column) across the training process. To quantify this relationship, we train a ridge regression model (with a single scalar weight) to predict brain alignment scores from benchmark scores using 10-fold cross-validation. The average R-squared value across these folds serves as our metric for comparing the relationship between formal/functional linguistic competence and brain alignment. These R-squared values are shown in each panel of Figure 4. Finally, we perform a

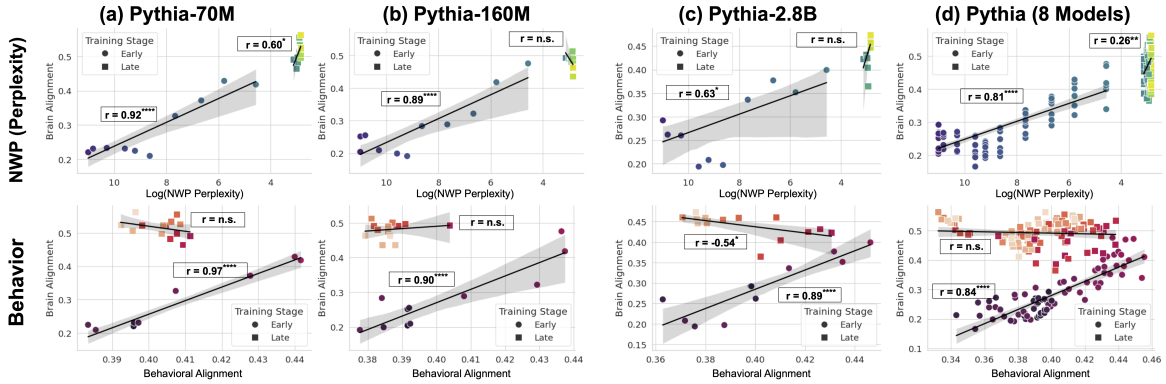

Figure 5: **NWP and Behavioral Alignment Correlate with Brain Alignment Only in Early Training. (Top Row)**: Correlation between brain alignment and language modeling loss shows a strong, significant relationship during early training (up to 2B tokens). While this correlation weakens in later stages (up to ~300B tokens). Results are shown for three models and the average of all 8 models (last column). **(Bottom Row)**: The same analysis, but for the correlation between brain alignment and behavioral alignment, revealing a similar trend—strong correlation early in training, but no significant relationship as models surpass human proficiency.

Wilcoxon signed-rank test on the distributions of R-squared values. This test reveals that formal linguistic competence is significantly more strongly correlated with brain alignment than functional competence ($W = 0.0$, $p < 0.002$). One possible explanation for why brain alignment emerges before formal linguistic competence is that existing LLM benchmarks assess performance using discrete accuracy thresholds (hard metrics), rather than capturing the gradual progression of competence through more nuanced, continuous measures (soft metrics) (Schaeffer et al., 2023). We show the individual benchmark scores across all checkpoints in Figure 8 in Appendix E.

5.3 LLMs Lose Behavioral Alignment

Do language models that improve in next-word prediction remain aligned with human behavioral and neural responses, or do they diverge as they surpass human proficiency? To answer this question we use the FUTRELL2018 benchmark, which has been widely used in previous research to measure linguistic behavior (Futrell et al., 2018; Schrimpf et al., 2021; Aw et al., 2023). This dataset consists of self-paced reading times for naturalistic story materials from 180 participants. Per-word reading times provide a measure of incremental comprehension difficulty, a cornerstone of psycholinguistic research for testing theories of sentence comprehension (Gibson, 1998; Smith and Levy, 2013; Brothers and Kuperberg, 2021; Shain et al., 2024). We measure alignment by calculating the Pearson correlation between a model's cross-entropy loss for a specific token in the sequence and the average human per-word reading time. The loss for words that comprise multiple tokens is added together before computing the correlation.

Early in training, LLMs align with this pattern, but as they surpass human proficiency (Shlegeris et al., 2022), their perplexity drops and they begin encoding statistical regularities that diverge from human intuition (Oh and Schuler, 2023; Steuer et al., 2023). This shift correlates with a decline in behavioral alignment, suggesting that superhuman models rely on different mechanisms than those underlying human language comprehension. Figure 5 shows that brain alignment initially correlates with perplexity and behavioral alignment, but only during the early stages of training (up to ~2B tokens). Beyond this point, these correlations diminish. In larger models, we observe a negative correlation between brain alignment and behavioral alignment in the later stages of training. This trend reinforces that early training aligns LLMs with human-like processing as also observed in earlier stages, while in later stages their language mechanisms diverge from humans.

6 Conclusion

In this work, we investigate how brain alignment in LLMs evolves throughout training, revealing different learning processes at play. We demonstrate that alignment with the human language network (LN) primarily correlates with formal linguistic competence (Mahowald et al., 2024), peaking and saturating early in training. In contrast, func-

tional linguistic competence, which involves world knowledge and reasoning, continues to grow beyond this stage. These findings suggest that the LN primarily encodes syntactic and compositional structure, in line with the literature of language neuroscience (Fedorenko et al., 2024a), while broader linguistic functions may rely on other cognitive systems beyond the LN. This developmental approach reveals when brain-like representations emerge, offering a dynamic perspective compared to prior work focused on fully trained models. For example, Oota et al. (2023) demonstrated that syntactic structure contributes to alignment by selectively removing specific properties from already trained models. In contrast, we show that formal linguistic competence actively drives brain alignment during the early phases of training. Similarly, Hosseini et al. (2024) reported that models achieve strong alignment with limited data; we identify why: the brain-like representations emerge as soon as core formal linguistic knowledge is acquired. Further, their study evaluated only four training checkpoints and 2 models on a single dataset (PEREIRA2018). Our study evaluated eight models (14M–6.7B parameters) across 34 checkpoints spanning 300B tokens, and used five neural benchmarks within a rigorous brain-scoring framework. This extensive design enabled fine-grained correlations with both formal and functional linguistic benchmarks and ensured our results are robust and generalizable.

We also show that model size is not a reliable predictor of brain alignment when controlling for the number of features (see Appendix I). Instead, alignment is shaped by architectural inductive biases, token integration mechanisms, and training dynamics. Our standardized brain-scoring framework eliminates contextualization biases from previous work, ensuring more rigorous evaluations. Finally, we demonstrate that current brain alignment benchmarks are not saturated, indicating that LLMs can still be improved in modeling human language processing. Together, these findings challenge prior assumptions about how alignment emerges in LLMs and provide new insights into the relationship between artificial and biological language processing.

Limitations

While this study offers a comprehensive analysis of brain alignment in LLMs, several open questions remain. If functional competence extends beyond the language network, future work should explore which additional brain regions LLMs align with as they develop reasoning and world knowledge, particularly in other cognitive networks like the multiple demand (Duncan and Owen, 2000) or theory of mind network (Saxe and Kanwisher, 2003; Saxe and Powell, 2006). Our findings suggest that LLM brain alignment studies should be broadened from the LN to downstream representations underlying other parts of cognition. This raises the question of whether specific transformer units specialize in formal vs. functional linguistic competence (AlKhamissi et al., 2025).

One other limitation of our study is that we rely exclusively on brain data collected from experiments conducted with English stimuli. As such, we do not explore whether our findings generalize across languages. This remains an open question and warrants further investigation. That said, evidence from cross-linguistic neuroscience research studying 45 languages from 12 language families (Malik-Moraleda et al., 2022) suggests the existence of a universal language network in the brain that is robust across languages and language families, both in topography and core functional properties.

Finally, a key question remains: Does LLM alignment evolution mirror human language acquisition? Comparing LLM representations to developmental data could reveal insights into learning trajectories and help differentiate formal from functional language learning. Expanding brain-scoring benchmarks and incorporating multimodal models will help address these questions, further bridging the gap between artificial and biological intelligence and deepening our understanding of how both systems process and represent language.

Ethical Statement

This research relies on previously published neuroimaging (fMRI, ECoG) and behavioral datasets, collected by the original research groups under their institutional ethical guidelines with informed consent and IRB/ethics approval. Our work involved only secondary analysis of de-identified data, with no new data collection or direct participant interaction, and we remain committed to using such data responsibly and respectfully.

Acknowledgments

We thank the members of the EPFL NeuroAI and NLP labs for their valuable feedback and insightful

suggestions. We also gratefully acknowledge the support of the Swiss National Science Foundation (No. 215390), Innosuisse (PFFS-21-29), the EPFL Center for Imaging, Sony Group Corporation, and a Meta LLM Evaluation Research Grant.

References

Badr AlKhamissi, Millicent Li, Asli Celikyilmaz, Mona T. Diab, and Marjan Ghazvininejad. 2022. A review on language models as knowledge bases. *ArXiv*, abs/2204.06031.

Badr AlKhamissi, Greta Tuckute, Antoine Bosselut, and Martin Schrimpf. 2025. The LLM language network: A neuroscientific approach for identifying causally task-relevant units. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 10887–10911, Albuquerque, New Mexico. Association for Computational Linguistics.

Loubna Ben Allal, Anton Lozhkov, Elie Bakouch, Gabriel Martín Blázquez, Guilherme Penedo, Lewis Tunstall, Andrés Marafioti, Hynek Kydlíček, Agustín Piqueres Lajarín, Vaibhav Srivastav, and 1 others. 2025. Smollm2: When smol goes big–data-centric training of a small language model. *arXiv preprint arXiv:2502.02737*.

Khai Loong Aw, Syrielle Montariol, Badr AlKhamissi, Martin Schrimpf, and Antoine Bosselut. 2023. Instruction-tuning aligns llms to the human brain.

Elizabeth Bates, Stephen M. Wilson, Ayse Pinar Saygin, Frederic Dick, Martin I. Sereno, Robert T. Knight, and Nina F. Dronkers. 2003. Voxel-based lesion–symptom mapping. *Nature Neuroscience*, 6(5):448–450.

Yael Benn, Iain D. Wilkinson, Ying Zheng, Kathrin Cohen Kadosh, Charles A.J. Romanowski, Michael Siegal, and Rosemary Varley. 2013. Differentiating core and co-opted mechanisms in calculation: The neuroimaging of calculation in aphasia. *Brain and Cognition*, 82(3):254–264.

Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar Van Der Wal. 2023. Pythia: a suite for analyzing large language models across training and scaling. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Jeffrey R. Binder, Julie A. Frost, Thomas A. Hammeke, Robert W. Cox, Stephen M. Rao, and Thomas Prieto. 1997. Human brain language areas identified by functional magnetic resonance imaging. *The Journal of Neuroscience*, 17(1):353–362.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2019. Piqa: Reasoning about physical commonsense in natural language. In *AAAI Conference on Artificial Intelligence*.

Idan Blank, Nancy Kanwisher, and Evelina Fedorenko. 2014. A functional dissociation between language and multiple-demand systems revealed in patterns of BOLD signal fluctuations. *Journal of Neurophysiology*, 112(5):1105–1118.

Trevor Brothers and Gina R Kuperberg. 2021. Word predictability effects are linear, not logarithmic: Implications for probabilistic models of sentence comprehension. *Journal of Memory and Language*, 116:104174.

Santiago A Cadena, George H Denfield, Edgar Y Walker, Leon A Gatys, Andreas S Tolias, Matthias Bethge, and Alexander S Ecker. 2019. Deep convolutional models improve predictions of macaque v1 responses to natural images. *PLoS computational biology*, 15(4):e1006897.

Charlotte Caucheteux and Jean-Rémi King. 2022. Brains and algorithms partially converge in natural language processing. *Communications biology*, 5(1):134.

Xuanyi Chen, Josef Affourtit, Rachel Ryskin, Tamar I Regev, Samuel Norman-Haignere, Olessia Jouravlev, Saima Malik-Moraleda, Hope Kean, Rosemary Varley, and Evelina Fedorenko. 2023. The human language system, including its inferior frontal component in "broca's area," does not support music perception. *Cerebral Cortex*, 33(12):7904–7929.

Radoslaw Martin Cichy, Aditya Khosla, Dimitrios Pantazis, Antonio Torralba, and Aude Oliva. 2016. Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific reports*, 6(1):27755.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *ArXiv*, abs/1803.05457.

John Duncan and Adrian M Owen. 2000. Common regions of the human frontal lobe recruited by diverse cognitive demands. *Trends in Neurosciences*, 23(10):475–483.

Jenelle Feather, Meenakshi Khosla, N. Apurva, Ratan Murty, and Aran Nayebi. 2025. Brain-model evaluations need the neuroai turing test.

Evelina Fedorenko. 2014. The role of domain-general cognitive control in language comprehension. *Frontiers in Psychology*, 5.

Evelina Fedorenko, Michael K Behr, and Nancy Kanwisher. 2011. Functional specificity for high-level linguistic processing in the human brain. *Proceedings of the National Academy of Sciences*, 108(39):16428–16433.

Evelina Fedorenko, Po-Jang Hsieh, Alfonso Nieto-Castanon, Susan L. Whitfield-Gabrieli, and Nancy G. Kanwisher. 2010. New method for fmri investigations of language: defining rois functionally in individual subjects. *Journal of neurophysiology*, 104 2:1177–94.

Evelina Fedorenko, Anna A. Ivanova, and Tamar I. Regev. 2024a. The language network as a natural kind within the broader landscape of the human brain. *Nature Reviews Neuroscience*, 25(5):289–312.

Evelina Fedorenko, Josh H. McDermott, Sam Norman-Haignere, and Nancy Kanwisher. 2012. Sensitivity to musical structure in the human brain. *Journal of Neurophysiology*, 108(12):3289–3300.

Evelina Fedorenko, Steven T. Piantadosi, and Edward A. F. Gibson. 2024b. Language is primarily a tool for communication rather than thought. *Nature*, 630(8017):575–586.

Evelina Fedorenko, Terri L. Scott, Peter Brunner, William G. Coon, Brianna Pritchett, Gerwin Schalk, and Nancy Kanwisher. 2016. Neural correlate of the construction of sentence meaning. *Proceedings of the National Academy of Sciences*, 113(41):E6256–E6262.

Ebrahim Feghhi, Nima Hadidi, Bryan Song, Idan A. Blank, and Jonathan C. Kao. 2024. What are large language models mapping to in the brain? a case against over-reliance on brain scores.

Richard Futrell, Edward Gibson, Harry J. Tily, Idan Blank, Anastasia Vishnevetsky, Steven Piantadosi, and Evelina Fedorenko. 2018. The natural stories corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, and 5 others. 2024. A framework for few-shot language model evaluation.

Jon Gauthier, Jennifer Hu, Ethan Wilcox, Peng Qian, and Roger Levy. 2020. SyntaxGym: An online platform for targeted evaluation of language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 70–76, Online. Association for Computational Linguistics.

Franziska Geiger, Martin Schrimpf, Tiago Marques, and James J DiCarlo. 2022. Wiring up vision: Minimizing supervised synaptic updates needed to produce a primate ventral stream. In *International Conference on Learning Representations 2022 Spotlight*.

Edward Gibson. 1998. Linguistic complexity: locality of syntactic dependencies. *Cognition*, 68(1):1–76.

Ariel Goldstein, Zaid Zada, Eliav Buchnik, Mariano Schain, Amy Price, Bobbi Aubrey, Samuel A. Nastase, Amir Feder, Dotan Emanuel, Alon Cohen, Aren Jansen, Harshvardhan Gazula, Gina Choe, Aditi Rao, Catherine Kim, Colton Casto, Lora Fanda, Werner Doyle, Daniel Friedman, and 13 others. 2022. Shared computational principles for language processing in humans and deep language models. *Nature Neuroscience*, 25(3):369–380.

Maria Luisa Gorno-Tempini, Nina F. Dronkers, Katherine P. Rankin, Jennifer M. Ogar, La Phengrasamy, Howard J. Rosen, Julene K. Johnson, Michael W. Weiner, and Bruce L. Miller. 2004. Cognition and anatomy in three variants of primary progressive aphasia. *Annals of Neurology*, 55(3):335–346.

Peter Hagoort. 2019. The neurobiology of language beyond single-word processing. *Science*, 366(6461):55–58.

Sarah E Harvey, Brett W. Larsen, and Alex H Williams. 2023. Duality of bures and shape distances with implications for comparing neural representations. In *UniReps: the First Workshop on Unifying Representations in Neural Models*.

Eghbal A Hosseini, Martin Schrimpf, Yian Zhang, Samuel Bowman, Noga Zaslavsky, and Evelina Fedorenko. 2024. Artificial neural network language models predict human brain responses to language even after a developmentally realistic amount of training. *Neurobiology of Language*, pages 1–21.

Jennifer Hu, Hannah Small, Hope Kean, Atsushi Takahashi, Leo Zekelman, Daniel Kleinman, Elizabeth Ryan, Alfonso Nieto-Castañón, Victor Ferreira, and Evelina Fedorenko. 2023. Precision fmri reveals that the language-selective network supports both phrase-structure building and lexical access during language production. *Cerebral Cortex*, 33(8):4384–4404.

Jie Huang and Kevin Chen-Chuan Chang. 2023. Towards reasoning in large language models: A survey. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1049–1065, Toronto, Canada. Association for Computational Linguistics.

Carina Kauf, Greta Tuckute, Roger Levy, Jacob Andreas, and Evelina Fedorenko. 2023. Lexical-Semantic Content, Not Syntactic Structure, Is the Main Contributor to ANN-Brain Similarity of fMRI Responses in the Language Network. *Neurobiology of Language*, pages 1–36.

Atlas Kazemian, Eric Elmoznino, and Michael F. Bonner. 2024. Convolutional architectures are cortex-aligned de novo. *bioRxiv*.

Alexander JE Kell, Daniel LK Yamins, Erica N Shook, Sam V Norman-Haignere, and Josh H McDermott. 2018. A task-optimized neural network replicates human auditory behavior, predicts brain responses,

and reveals a cortical processing hierarchy. *Neuron*, 98(3):630–644.

Seyed Mahdi Khaligh-Razavi and Nikolaus Kriegeskorte. 2014. Deep Supervised, but Not Unsupervised, Models May Explain IT Cortical Representation. *PLoS Computational Biology*, 10(11). Publisher: Public Library of Science ISBN: 1553-7358 (Electronic)\r1553-734X (Linking).

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. 2019. Similarity of neural network representations revisited. In *International conference on machine learning*, pages 3519–3529. PMLR.

Takuya Koumura, Hiroki Terashima, and Shigeto Furukawa. 2023. Human-like modulation sensitivity emerging through optimization to natural sound recognition. *Journal of Neuroscience*, 43(21):3876–3894.

Nikolaus Kriegeskorte, Marieke Mur, and Peter Bandettini. 2008. Representational similarity analysis - connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2.

Jonas Kubilius, Martin Schrimpf, Kohitij Kar, Rishi Rajalingham, Ha Hong, Najib Majaj, Elias Issa, Pouya Bashivan, Jonathan Prescott-Roy, Kailyn Schmidt, Aran Nayebi, Daniel Bear, Daniel L Yamins, and James J DiCarlo. 2019. Brain-like object recognition with high-performing shallow recurrent anns. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Benjamin Lipkin, Greta Tuckute, Josef Affourtit, Hannah Small, Zachary Mineroff, Hope Kean, Olessia Jouravlev, Lara Rakocevic, Brianna Pritchett, Matthew Siegelman, Caitlyn Hoeflin, Alvincé Pongos, Idan A. Blank, Melissa Kline Struhl, Anna Ivanova, Steven Shannon, Aalok Sathe, Malte Hoffmann, Alfonso Nieto-Castañón, and Evelina Fedorenko. 2022. Probabilistic atlas for the language network based on precision fmri data from>800 individuals. *Scientific Data*, 9(1).

Kyle Mahowald, Anna A Ivanova, Idan A Blank, Nancy Kanwisher, Joshua B Tenenbaum, and Evelina Fedorenko. 2024. Dissociating language and thought in large language models. *Trends in Cognitive Sciences*.

Saima Malik-Moraleda, Dima Ayyash, Jeanne Gallée, Josef Affourtit, Malte Hoffmann, Zachary Mineroff, Olessia Jouravlev, and Evelina Fedorenko. 2022. An investigation across 45 languages and 12 language families reveals a universal language network. *Nature Neuroscience*, 25(8):1014–1019.

Juliette Millet and Jean-Rémi King. 2021. Inductive biases, pretraining and fine-tuning jointly account for brain responses to speech. *ArXiv*, abs/2103.01032.

Martin M Monti, Lawrence M Parsons, and Daniel N Osherson. 2012. Thought beyond language: neural dissociation of algebra and natural language. *Psychological science*, 23(8):914–922.

Samuel A. Nastase, Yun-Fei Liu, Hanna Hillman, Asieh Zadbood, Liat Hasenfratz, Neggin Keshavarzian, Janice Chen, Christopher J. Honey, Yaara Yeshurun, Mor Regev, and et al. 2021. The "narratives" fmri dataset for evaluating models of naturalistic language comprehension. *Scientific Data*, 8(1).

Byung-Doh Oh and William Schuler. 2023. Why does surprisal from larger transformer-based language models provide a poorer fit to human reading times? *Transactions of the Association for Computational Linguistics*, 11:336–350.

Subba Reddy Oota, Manish Gupta, and Mariya Toneva. 2023. Joint processing of linguistic properties in brains and language models. *Preprint*, arXiv:2212.08094.

Alexandre Pasquiou, Yair Lakretz, John Hale, Bertrand Thirion, and Christophe Pallier. 2022. Neural language models are not born equal to fit brain data, but training helps. *Preprint*, arXiv:2207.03380.

Guilherme Penedo, Hynek Kydlíček, Loubna Ben allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. 2024. The fineweb datasets: Decanting the web for the finest text data at scale. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Francisco Pereira, Bin Lou, Brianna Pritchett, Samuel Ritter, Samuel J. Gershman, Nancy Kanwisher, Matthew Botvinick, and Evelina Fedorenko. 2018. Toward a universal decoder of linguistic meaning from brain activation. *Nature Communications*, 9(1):963.

Cathy J. Price. 2010. The anatomy of language: a review of 100 fmri studies published in 2009. *Annals of the New York Academy of Sciences*, 1191(1):62–88.

Neil Rathi, Johannes Mehrer, Badr AlKhamissi, Taha Binhuraib, Nicholas M. Blauch, and Martin Schrimpf. 2025. TopoLM: Brain-like spatio-functional organization in a topographic language model. In *International Conference on Learning Representations (ICLR)*.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Winogrande. *Communications of the ACM*, 64:99 – 106.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019. Social IQa: Commonsense reasoning about social interactions. 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 4463–4473, Hong Kong, China. Association for Computational Linguistics.

R Saxe and N Kanwisher. 2003. People thinking about thinking peoplethe role of the temporoparietal junction in "theory of mind". *NeuroImage*, 19(4):1835–1842.

Rebecca Saxe, Matthew Brett, and Nancy Kanwisher. 2006. Divide and conquer: a defense of functional localizers. *Neuroimage*, 30(4):1088–1096.

Rebecca Saxe and Lindsey J. Powell. 2006. It's the thought that counts: Specific brain regions for one component of theory of mind. *Psychological Science*, 17(8):692–699.

Rylan Schaeffer, Brando Miranda, and Oluwasanmi Koyejo. 2023. Are emergent abilities of large language models a mirage? *ArXiv*, abs/2304.15004.

Martin Schrimpf, Idan Asher Blank, Greta Tuckute, Carina Kauf, Eghbal A. Hosseini, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. 2021. The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, 118(45):e2105646118.

Martin Schrimpf, Jonas Kubilius, Ha Hong, Najib J. Majaj, Rishi Rajalingham, Elias B. Issa, Kohitij Kar, Pouya Bashivan, Jonathan Prescott-Roy, Franziska Geiger, Kailyn Schmidt, Daniel L. K. Yamins, and James J. DiCarlo. 2018. Brain-Score: Which Artificial Neural Network for Object Recognition is most Brain-Like? preprint, Neuroscience.

Martin Schrimpf, Jonas Kubilius, Michael J. Lee, N. Apurva Ratan Murty, Robert Ajemian, and James J. DiCarlo. 2020. Integrative benchmarking to advance neurally mechanistic models of human intelligence. *Neuron*, 108(3):413–423.

Cory Shain, Clara Meister, Tiago Pimentel, Ryan Cotterell, and Roger Levy. 2024. Large-scale evidence for logarithmic effects of word predictability on reading time. *Proceedings of the National Academy of Sciences*, 121(10):e2307876121.

Buck Shlegeris, Fabien Roger, Lawrence Chan, and Euan McLean. 2022. Language models are better than humans at next-token prediction. *ArXiv*, abs/2212.11281.

Michael Siegal and Rosemary Varley. 2006. Aphasia, language, and theory of mind. *Social Neuroscience*, 1(3–4):167–174.

Nathaniel J. Smith and Roger Levy. 2013. The effect of word predictability on reading time is logarithmic. *Cognition*, 128(3):302–319.

Julius Steuer, Marius Mosbach, and Dietrich Klakow. 2023. Large gpt-like models are bad babies: A closer look at the relationship between linguistic competence and psycholinguistic measures. *arXiv preprint arXiv:2311.04547*.

Damien Teney, Armand Nicolicioiu, Valentin Hartmann, and Ehsan Abbasnejad. 2024. Neural redshift: Random networks are not random functions. *Preprint*, arXiv:2403.02241.

Greta Tuckute, Jenelle Feather, Dana Boebinger, and Josh H. McDermott. 2023. Many but not all deep neural network audio models capture brain responses and exhibit correspondence between model stages and brain regions. *PLOS Biology*, 21(12):1–70.

Greta Tuckute, Nancy Kanwisher, and Evelina Fedorenko. 2024a. Language in brains, minds, and machines. *Annual Review of Neuroscience*, 47.

Greta Tuckute, Aalok Sathe, Shashank Srikant, Maya Taliaferro, Mingye Wang, Martin Schrimpf, Kendrick Kay, and Evelina Fedorenko. 2024b. Driving and suppressing the human language network using large language models. *Nature Human Behaviour*, pages 1–18.

Rosemary Varley and Michael Siegal. 2000. Evidence for cognition without grammar from causal reasoning and 'theory of mind' in an agrammatic aphasic patient. *Current Biology*, 10(12):723–726.

Rosemary A. Varley, Nicolai J. C. Klessinger, Charles A. J. Romanowski, and Michael Siegal. 2005. Agrammatic but numerate. *Proceedings of the National Academy of Sciences*, 102(9):3519–3524.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2019. Blimp: The benchmark of linguistic minimal pairs for english. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Daniel LK Yamins, Ha Hong, Charles F Cadieu, Ethan A Solomon, Darren Seibert, and James J DiCarlo. 2014. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the national academy of sciences*, 111(23):8619–8624.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? In *Annual Meeting of the Association for Computational Linguistics*.

Chengxu Zhuang, Siming Yan, Aran Nayebi, Martin Schrimpf, Michael C. Frank, James J. DiCarlo, and Daniel L.K. Yamins. 2021. Unsupervised neural network models of the ventral visual stream. *Proceedings of the National Academy of Sciences (PNAS)*, 118(3). Publisher: Cold Spring Harbor Laboratory.

Appendix

A Neuroimaging & Behavioral Datasets

Table 1 shows the different neuroimaging and behavioral datasets used in this work, along with the dataset modality, presentation mode, and a stimulus example.

A.1 Neuroimaging Datasets

(Pereira et al., 2018) This dataset consists of fMRI activations (blood-oxygen-level-dependent; BOLD responses) recorded as participants read short passages presented one sentence at a time for 4 s. The dataset is composed of two distinct experiments: one with 9 subjects presented with 384 sentences, and another with 6 subjects presented with 243 sentences each. The passages in each experiment spanned 24 different topics. The results reported for this dataset are the average alignment across both experiments after normalizing with their respective cross-subject consistency estimates.

(Blank et al., 2014) This dataset also involves fMRI signals but recorded from only 12 functional regions of interest (fROI) instead of the higher resolution signal used by Pereira et al. (2018). The data was collected from 5 participants as they listened to 8 long naturalistic stories that were adapted from existing fairy tales and short stories (Futrell et al., 2018). Each story was approximately 5 minutes long, averaging up to 165 sentences, providing a much longer context length than the other neuroimaging datasets. When measuring brain alignment, we use the input stimuli of the last 32 TRs as the model's context.

(Fedorenko et al., 2016) This dataset captures ECoG signals from 5 participants as they read 8-word-long sentences presented one word at a time for 450 or 700 ms. Following (Schrimpf et al., 2021) we select the 52/80 sentences that were presented to all participants.

(Tuckute et al., 2024b) In this dataset, 5 participants read 1000 6-word sentences presented one sentence at a time for 2 s. BOLD responses from voxels in the language network were averaged within each participant and then across participants to yield an overall average language network response to each sentence. The stimuli used span a large part of the linguistic space, enabling model-brain comparisons across a wide range of single sentences. Sentence presentation order was randomized across participants. In combination with the diversity in linguistic materials, this dataset presents a particularly challenging dataset for model evaluation.

Narratives Dataset (Nastase et al., 2021) This dataset consists of fMRI data collected while human subjects listened to 27 diverse spoken story stimuli. The collection includes 345 subjects, 891 functional scans, and approximately 4.6 hours of unique audio stimuli. For our story-based analysis, we focused on 5 participants who each listened to both the LUCY and TUNNEL stories. Since functional localization was not performed in the NARRATIVES dataset, we approximated language regions by extracting the top-10% voxels from each anatomically defined language region according to a probabilistic atlas for the human language system (Lipkin et al., 2022). Due to the limited corpus of two stories, traditional 10-fold cross-validation was not feasible. To implement topic-based splitting while maintaining methodological rigor, we partitioned each story into n distinct segments, with each segment functioning as an independent narrative unit. This segmentation approach effectively prevented cross-contamination of contextual information between splits, thereby preserving the integrity of our evaluation framework.

A.2 Behavioral Dataset

(Futrell et al., 2018) This dataset consists of self-paced reading times for each word from 180 participants. The stimuli include 10 stories from the Natural Stories Corpus (Futrell et al., 2018), similar to BLANK2014. Each participant read between 5 and all 10 stories.

B Rigorous Brain-Scoring

Despite progress in linking LLMs to neural activity, there's no standard for comparing brain alignment across datasets and conditions. Here, we aim to establish a set of desiderata for evaluating brain alignment. For a model to be considered truly brain-aligned, two key criteria must be met. First, high alignment scores should indicate that the model captures stimulus-driven responses—meaning that when presented with a random sequence of tokens, alignment should drop significantly compared to original linguistic stimuli. Second, a brain-aligned model should generalize effectively to new linguistic contexts rather than overfitting to specific ex-

Dataset	Modality	Presentation	Stimulus Example
PEREIRA2018	fMRI	Reading	Accordions produce sound with bellows ...
BLANK2014	fMRI	Listening	A clear and joyous day it was and out on the wide ...
FEDORENKO2016	ECoG	Reading	'ALEX', 'WAS', 'TIRED', 'SO', 'HE', 'TOOK', ...
TUCKUTE2024	fMRI	Reading	The judge spoke, breaking the silence.
NARRATIVES	fMRI	Listening	Okay so getting back to our story about uh Lucy ...
FUTRELL2018	Reading Times	Reading	A clear and joyous day it was and out on the wide ...

Table 1: **Datasets Used for Evaluating Model Alignment.** Neuroimaging datasets were collected via either functional magnetic resonance imaging (fMRI) or electrocorticography (ECoG). Stimuli range from short sentences (FEDORENKO2016, TUCKUTE2024) to paragraphs (PEREIRA2018) and entire stories (BLANK2014, NARRATIVES, FUTRELL2018) and were presented either visually or auditorily. FUTRELL2018 is a behavioral dataset.

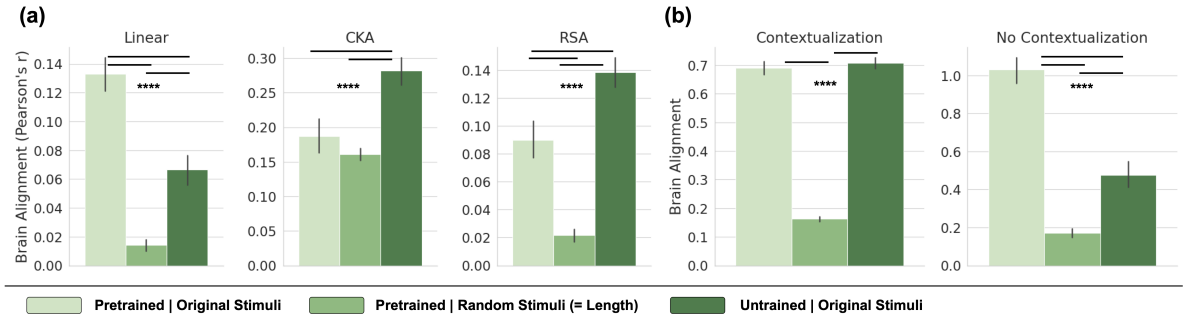

Figure 6: **Evaluating Brain Alignment with Linear Predictivity and No Contextualization is Most Stringent**. **(a)** Average brain alignment across 8 Pythia models under three conditions: (1) a pretrained model processing the original stimuli, (2) a pretrained model processing *random* sequences of the same length (averaged over five random seeds) as a control condition, and (3) the model with untrained parameters processing the original stimuli. The linear predictivity metric differentiates between meaningful and random stimuli most strongly, while RSA and CKA overestimate alignment. **(b)** Brain alignment on the PEREIRA2018 dataset under two cross-validation schemes: with contextualization (random sentence split) and without contextualization (story-based split).

amples. We address these two points in Section 4 to justify our choice of metric and cross-validation scheme for each dataset (see Figure 6). For all benchmarks, we localize language-selective units, which is consistent with neural site selection in neuroscience experiments and allows for fair comparisons across models irrespective of model size (AlKhamissi et al., 2025). A key limitation of previous methods is their reliance on the raw hidden state dimensions, which inherently favors larger models by providing a greater feature space and artificially inflating alignment scores.

C Brain-Score Using Additional Metrics

Centered Kernel Alignment (CKA) Kornblith et al. (2019) introduced CKA as a substitute for Canonical Correlation Analysis (CCA) to assess the similarity between neural network representations. Unlike linear predictivity, it is a non-parameteric metric and therefore does not require any additional training. CKA is particularly effective with high-dimensional representations, and its reliability in identifying correspondences between representations in networks trained from different initializations (Kornblith et al., 2019).

Representational Similarity Analysis (RSA) (Kriegeskorte et al., 2008) introduced RDMs as a solution to the challenge of integrating brain-activity measurements, behavioral observations, and computational models in systems neuroscience. RDMs are part of a broader analytical framework referred to as representational similarity analysis (RSA). In practical terms, to compute the dissimilarity matrix for an N-dimensional network's responses to M different stimuli, an $M \times M$ matrix of distances between all pairs of evoked responses is generated for both brain activity and the language model's activations (Harvey et al., 2023). The cor-

Num Tokens	Pereira2018	Blank2014	Tuckute2024	Fedorenko2016	Narratives	Avg
250B	1.00	0.19	0.47	0.78	0.04	0.50
500B	0.97	0.08	0.51	0.87	0.04	0.49
750B	0.99	0.08	0.52	0.78	0.04	0.48
1T	1.07	0.12	0.55	0.84	0.04	0.52
1.25T	1.00	0.12	0.50	0.82	0.03	0.49
1.5T	1.00	0.12	0.52	0.79	0.03	0.49
1.75T	0.96	0.13	0.48	0.79	0.04	0.48
2T	1.05	0.15	0.56	0.84	0.04	0.53
2.25T	1.08	0.16	0.55	0.75	0.04	0.51
2.5T	1.12	0.17	0.52	0.72	0.01	0.51
2.75T	1.13	0.12	0.49	0.75	0.04	0.49
3T	1.03	0.26	0.51	0.55	0.01	0.47
3.25T	1.02	0.13	0.52	0.68	0.02	0.47
3.5T	1.04	0.14	0.52	0.72	0.04	0.49
3.75T	1.14	0.06	0.57	0.84	0.03	0.53
4T	1.05	0.13	0.63	0.82	0.05	0.54

Table 2: **Brain Alignment Performance of SmolLM2-360M Across Training Checkpoints.** Reported scores correspond to normalized correlations with neural responses from five benchmark datasets (Pereira2018, Blank2014, Tuckute2024, Fedorenko2016, Narratives), along with their average (Avg). These results assess the extent to which the model's internal representations align with activity in the human language network.

relation between these two matrices is then used as a measure of brain alignment.

D Brain Alignment Over Training

Figure 7 complements Figure 3 in the main paper, illustrating that brain alignment saturates early on in training for all models analyzed in this work.

E Formal & Functional Scores

Figure 8 presents the individual benchmark scores for both formal and functional linguistic competence across training. Formal benchmarks peak early, mirroring the trajectory of brain alignment, and remain saturated throughout training. In contrast, functional benchmarks continue to improve, reflecting the models' increasing ability to acquire factual knowledge and reasoning skills as they are trained on significantly more tokens using next-word prediction.

F Results on SmolLM2-360M

To assess the generalizability of our findings, we replicated our experiments using a model from a different language family. Specifically, we evaluated multiple training checkpoints of SmolLM2-360M on the brain alignment, formal, and functional linguistic competence benchmarks. Since SmolLM2 only provides checkpoints at intervals of 250B tokens, we cannot capture the gradual emergence of brain alignment and formal competence, both of which typically saturate around 4B–8B tokens. Given this limitation, our hypothesis was that brain alignment and formal competence would remain largely stable across these checkpoints, while functional competence would continue to improve. The results are consistent with this hypothesis as shown in Tables 2 and 3.

G Role of Weight Initialization

Figure 9 examines the effect of weight initialization variance on brain alignment in untrained models. We systematically vary the initialization standard deviation (sd) and find that the default HuggingFace (Wolf et al., 2019) initialization (sd = 0.02) achieves the highest alignment across datasets. This suggests that even before training begins, the choice of initialization can significantly influence how well a model's representations align with neural activity. This finding raises an intriguing hypothesis: could brain alignment, a computationally inexpensive metric, serve as a useful heuristic for selecting optimal initialization parameters? If so, it could help models learn tasks more efficiently and converge faster, reducing the need for extensive

Num Tokens	BLiMP	SyntaxGym	Avg (Formal)	ARC-Easy	ARC-Challenge	Social-IQA	PIQA	WinoGrande	HellaSwag	Avg (Functional)
250B	0.81	0.80	0.81	0.33	0.66	0.35	0.70	0.55	0.47	0.52
500B	0.80	0.78	0.79	0.78	0.66	0.35	0.70	0.56	0.49	0.53
750B	0.80	0.82	0.81	0.69	0.69	0.34	0.71	0.57	0.50	0.53
1T	0.81	0.78	0.80	0.69	0.69	0.35	0.71	0.57	0.50	0.54
1.25T	0.81	0.78	0.79	0.68	0.68	0.35	0.71	0.57	0.51	0.54
1.5T	0.81	0.80	0.80	0.69	0.68	0.35	0.72	0.56	0.51	0.54
1.75T	0.80	0.79	0.79	0.68	0.68	0.36	0.72	0.59	0.51	0.54
2T	0.81	0.81	0.81	0.69	0.69	0.35	0.72	0.59	0.52	0.54
2.25T	0.81	0.82	0.81	0.68	0.68	0.35	0.71	0.59	0.51	0.54
2.5T	0.81	0.82	0.82	0.68	0.68	0.36	0.70	0.56	0.52	0.54
2.75T	0.81	0.82	0.81	0.25	0.23	0.35	0.50	0.57	0.50	0.50
3T	0.81	0.81	0.81	0.25	0.23	0.35	0.50	0.57	0.50	0.50
3.25T	0.81	0.77	0.79	0.67	0.67	0.34	0.67	0.57	0.51	0.52
3.5T	0.81	0.79	0.80	0.71	0.71	0.38	0.72	0.58	0.53	0.55
3.75T	0.80	0.78	0.79	0.72	0.72	0.58	0.58	0.54	0.56	0.56
4T	0.81	0.79	0.80	0.73	0.73	0.39	0.74	0.61	0.56	0.57

Table 3: **Performance of SmolLM2-360M on Formal and Functional Linguistic Benchmarks Across Training Checkpoints.** Formal competence is measured using BLiMP and SyntaxGym (with averages reported as Avg Formal). Functional competence is measured using ARC-Easy, ARC-Challenge, Social-IQA, PIQA, WinoGrande, and HellaSwag (with averages reported as Avg Functional). Together, these results characterize the relationship between training progression and the development of different aspects of linguistic ability.

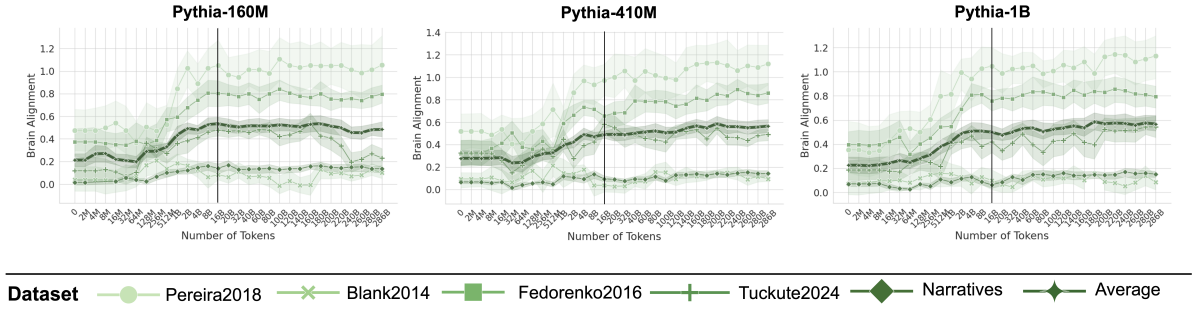

Figure 7: **Brain Alignment Saturates Early on in Training.** Plots complementing Figure 3 showing the brain alignment scores of three other models from the Pythia model suite with varying sizes (log x-axis up to 16B tokens, uneven spacing after black line). Scores are normalized by their cross-subject consistency scores. Alignment quickly peaks around 2–8B tokens before saturating or declining, regardless of model size.

trial-and-error in training from scratch. The results highlight the importance of architectural inductive biases and suggest that brain alignment may serve as a useful heuristic for optimizing model initialization.

H Effect of Number of Units on Brain Alignment

Figure 10 illustrates the impact of localizing more units on final brain alignment across the eight Pythia models used in this study. We find that increasing the number of units has minimal impact on the relative ranking of models, with only a slight increase in average alignment. Additionally, model size does not influence brain alignment once the number of units is controlled, reinforcing the idea that alignment is driven by feature selection rather than scale.

I Model Size Does Not Predict Alignment

Figure 12 presents the brain alignment for each dataset, along with the average alignment across datasets, for eight models of varying sizes from the Pythia model suite (final checkpoint). Contrary to the assumption that larger models exhibit higher brain alignment (Aw et al., 2023), we observe a decline in average alignment starting from 1B parameters up to 6.9B parameters, when controlling for feature size. This analysis is made possible by functional localization, which allows us to extract a fixed number of units from each model, rather than relying on hidden state dimensions, as done in previous studies. This approach ensures a fairer comparison among models. We show in Appendix H that increasing the number of localized units has minimal impact on the relative ranking of the models. Additionally, these findings align with expectations in the neuroscience language community, where it is widely believed that human language

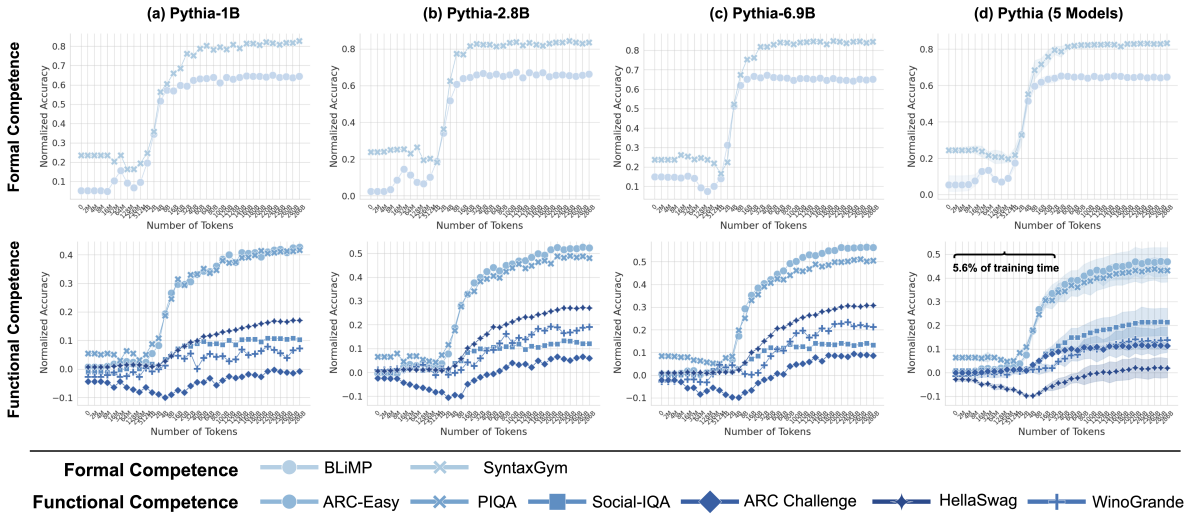

Figure 8: **Individual Benchmark Scores for Formal and Functional Competence. (a-c)**: each column shows the evolution of individual benchmark scores for formal competence (top) and functional competence (bottom) during training. Data is presented for Pythia models of three different sizes. **(d)**: the same as (a–c), with data averaged across models of five different sizes.

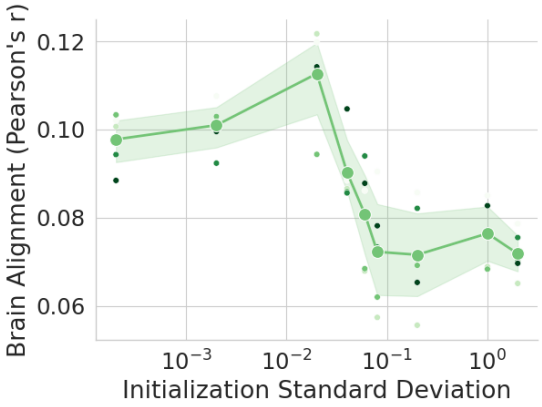

Figure 9: **Role of Weight Initialization on Brain Alignment in Untrained Models** The default initialization standard deviation in the HuggingFace library (sd = 0.02) yields the highest brain alignment for untrained models, suggesting that initialization choices play a crucial role in shaping alignment even before training begins.

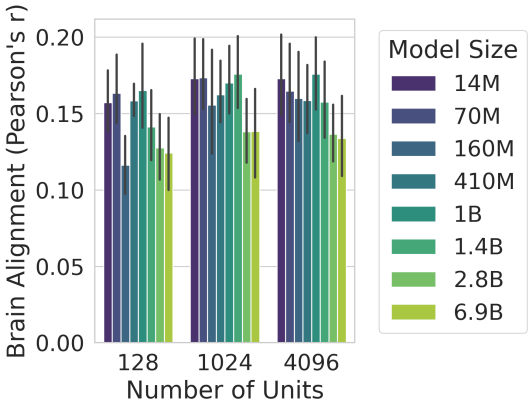

Figure 10: **The Effect of the Number of Localized Units on Final Brain Alignment** Brain alignment is evaluated after localizing 128, 1024, and 4096 units. While increasing the number of units slightly affects overall alignment, the relative ranking of models remains largely unchanged, indicating that model comparisons are robust to the choice of unit count.

processing does not require superhuman-scale models to capture neural activity in the brain's language network.

J Alignment with Other Brain Regions

As a control, we also examine alignment with non-language brain regions. Specifically, Figure 11 shows the brain alignment of three Pythia models with both the language network (LN) and V1—an early visual cortex region—on the PEREIRA2018 dataset. While alignment with the LN increases early in training (around 4B tokens) and then saturates, alignment with V1 remains largely unchanged throughout training. This divergence highlights a key aspect of LLM representations: they do not appear to encode low-level perceptual features, such as those processed in early visual areas. If models were learning perceptual structure from the stimuli, we would expect alignment with V1 to increase alongside LN alignment. Instead, the stability of V1 alignment across training suggests that language models selectively develop internal rep-

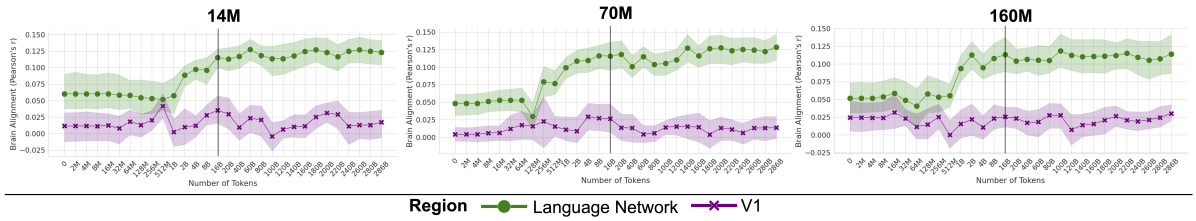

Figure 11: **Brain Alignment with the Language Network vs. V1 Across Training.** Raw brain alignment scores (Pearson's r) of three Pythia models of varying sizes are shown on the PEREIRA2018 dataset. The x-axis (log-scaled up to 16B tokens; then evenly spaced after the black line every 20B tokens) represents training progress. Alignment with V1, an early visual region, remains stable throughout training, while alignment with the language network (LN) increases around 4B tokens before plateauing.

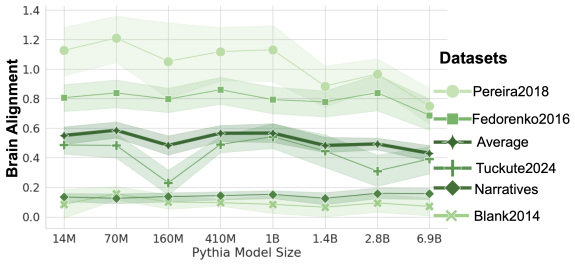

Figure 12: **Model Size Does Not Predict Brain Alignment when localizing a fixed set of language units**. Brain alignment across model sizes in the Pythia suite, measured at their final training checkpoints. Brain alignment is shown for each dataset, along with the average score across datasets, for eight models of varying sizes.

resentations that align with higher-order linguistic processing rather than general sensory processing.

One reason for not measuring alignment against other higher-level cognitive brain regions such as the default mode network (DMN), the multiple demand network (MD) or the theory of mind network (ToM) is due to a major limitation in current neuroimaging datasets: the linguistic stimuli used in studies with publicly available datasets (e.g., PEREIRA2018) do not reliably engage these higher-level cognitive regions, leading to substantial variability across individuals and thus much lower cross-subject consistency scores. Simply "looking" for alignment in the DMN or MD is therefore insufficient. Instead, we need new datasets that deliberately activate non-language networks and record item-level neural responses. For example, most MD studies rely on blocked fMRI designs (e.g., hard vs. easy math), yielding one activation estimate per condition rather than per stimulus. Such coarse measurements limit their utility to evaluate model-to-brain correspondence at the granularity of individual items. We expect alignment with the MD network, a brain region involved in logical reasoning, to track functional linguistic competence more than formal competence as models improve on relevant benchmarks. We leave this investigation for future work, pending the availability of suitable datasets.

K Cross-Subject Consistency Scores

Benchmark	Consistency Score
PEREIRA2018 (Exp 2)*	0.086
PEREIRA2018 (Exp 3)	0.144
BLANK2014	0.178
FEDORENKO2016	0.222
TUCKTUE2024	0.559
NARRATIVES	0.181
FUTRELL2018	0.858

Table 4: **Cross-Subject Consistency Scores** The values used to normalize the raw Pearson correlation. *PEREIRA2018 (Exp 2) was computed without extrapolation.

Table 4 shows the cross-subject consistency scores computed with extrapolation for the different benchmarks used in this work.