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 better modeled by formal than functional aspects of language.

Keywords: Language; Human Language Network; LLMs; Brain Alignment; Behavioral Alignment

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., 2024). 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., sentences 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; Tuckute et al., 2024; AlKhamissi et al., 2025).

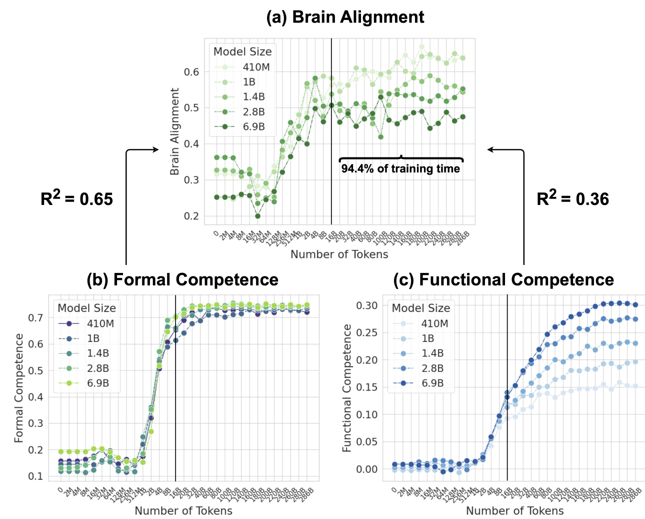

Figure 1: **Model Alignment with the Human Language Network is Primarily Driven by Formal than Functional Linguistic Competence. (a)** Average brain alignment across five Pythia models and five brain recording datasets, normalized by cross-subject consistency, throughout training. **(b)** Average normalized 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. The vertical black line is at 16B tokens.

Key Questions and Contributions

This work investigates four key questions, all aimed at distilling *why* LLM aligns to brain responses. Specifically, we investigate how linguistic competence emerges across training (developmental experience). We ask: (1) Is brain alignment primarily linked to formal or functional linguistic competence (Mahowald et al., 2024)? (2) Do language models diverge from humans as they surpass human-level prediction? (3) Do current LLMs fully account for the explained variance in brain alignment benchmarks? To answer these questions, we in-

* Equal Supervision

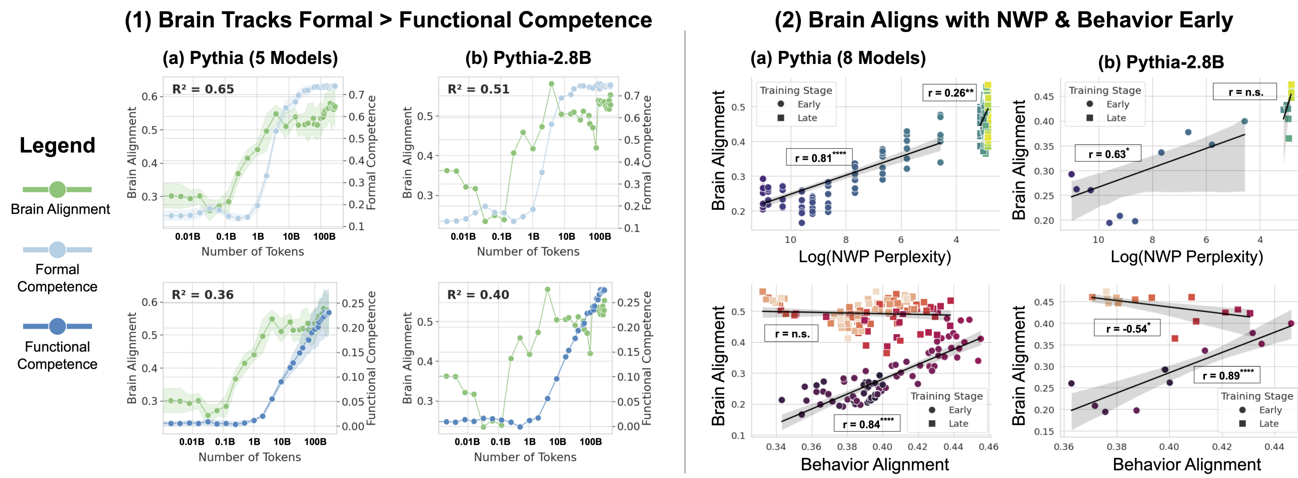

Figure 2: **(1) Formal Competence Tracks Brain Alignment More Closely Than Functional Competence.** Each column in (1) compares how the evolution of formal competence (top) and functional competence (bottom) tracks the evolution of brain alignment during training. The R^2 values quantify the strength 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)**: the same comparison as in (a), but with comparisons were made for PYTHIA 2.8B as an example. **(2) 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 the average of all 8 models (last column) the the 2.8B model . **(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.

troduce a rigorous brain-scoring framework to conduct a controlled and large-scale analysis of LLM brain alignment.

Results

Brain Alignment Over Training Figure 1(a) illustrates the brain alignment of 5 Pythia models across 5 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, 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 2.1 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 the average of five Pythia models and the 2.8B Pythia model across the training process. One possible explanation for why brain alignment emerges before formal linguistic competence is that existing LLM benchmarks assess performance using discrete accuracy thresholds, rather than capturing the gradual progression of competence through more nuanced, continuous measures (Schaeffer et al., 2023).

LLMs Lose Behavioral Alignment Human language processing is strongly modulated by prediction: unexpected words lead to longer reading times (Smith and Levy, 2013; Brothers and Kuperberg, 2021; Shain et al., 2024). 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 2.2 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.

References

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.

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.

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.

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.

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

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

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

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.

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*.

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.

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.

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.

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.

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*.

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