Language Evolution in Large Language Models and Humans: A Comparative Analysis of Developmental Linguistics Across Ages and Sensory Modalities

Gina Yu (gina.l.yu@vanderbilt.edu)
Neuroscience Department, Vanderbilt University

Jad El Harake (jad.el.harake@vanderbilt.edu)
Biomedical Engineering Department, Vanderbilt University

Stephen Chong Zhao (chong.zhao.1@vanderbilt.edu)
Data Science Institute, Vanderbilt University

Andrew Bender (abender@health.ucsd.edu)
Neurosciences Graduate Program, University of California San Diego

Jason Lee (jason.j.lee@vanderbilt.edu)
Computer Science Department, Vanderbilt University

Trisha Mazumdar (trisha.mazumdar@vanderbilt.edu)
Computer Science Department, Vanderbilt University

Adaline Leong (jia.yin.leong@vanderbilt.edu)
Computer Science Department, Vanderbilt University

Mark Wallace (mark.wallace@Vanderbilt.Edu)
Psychology Department, Vanderbilt University

David A. Tovar (david.tovar@vanderbilt.edu)
Psychology Department, Vanderbilt University

Large Language Models (LLMs) are able to adopt various roles, including simulating language use across different ages, which suggests an understanding of language evolution in humans. This study investigates the extent to which LLMs can mimic human developmental linguistics, comparing LLM interpretation of images and text to structural MRI data of the human brain at different ages. We analyze LLM output complexity through semantic embeddings and assess the embeddings' correlation with human brain development. Our findings indicate that the model's ability to replicate age-specific developmental stages varies significantly with the input modality. In addition, substantial correlations are observed between LLM output and structural brain changes in humans. We found that brain-network correlations to LLM output were specific to the input modality chosen. For example, LLM image interpretations were more closely tied to visual networks than were text interpretations. This work has implications for our understanding of LLMs as well as our understanding of the developmental linguistic and sensory changes that occur over a human development.

Keywords: Large Language Models; Semantic Embeddings; Age Development; MRI

Introduction

The rapid evolution of large language models (LLMs) is partly attributable to the ease of specific role prompting and requests to achieve intended results. However, short-comings have been reported in LLM performance, including hallucination artifacts and implicit biases, which can be difficult to detect (McKenna et al., 2023). A recent study on LLM in-context impersonation revealed that prompting the model to impersonate children recovers aspects of human-like developmental stages (Salewski et al., 2023). However, it remains unclear how accurately these developmental stages can be reproduced with LLMs compared to human behavior. Previous literature has shown LLM behavior and human brain function converges with good agreement between neural activity and embedding vectors (Caucheteux & King, 2022). Neural signals have been shown to correlate with LLM embeddings via representation similarity analysis (Li et al., 2023). Although functional imaging has been used to correlate neural activity with LLM embeddings, the relationship between LLMs and structural brain composition has not yet been evaluated. In this study, a framework is proposed for the evaluation of language evolution in LLM as a function of age. Anatomical brain imaging is investigated as a potential correlate to LLM embeddings. In addition, the effect of input modality is studied by considering text-only, image-only, and a combination of image and text.

Methods

Study Design. Short stories were generated with the use of GPT4 and DALL-E 3. Each story consisted of 3-5 images with text narratives. The GPT-4 API was then prompted to "Describe the story" based on the images only, the text only, and a multi-modal input of both. This process was repeated 30 times and assigning the LLM a different age as a role. The ages assigned ranged from 5-50 years old, sampling each year between ages 5 and 30 (5,6,7...30 years old) then sampling at 5 year-intervals (35,40,45,50 years old). The embedding vectors were obtained for each story, age, and input modality. Representational dissimilarity matrices (RDMs) were generated across ages for each story and each input modality. A dataset of structural imaging of the human brain at different ages (Bethlehem et al., n.d.) was used to generate RDMs across brain areas for comparison. Squared Euclidean distance was used for RDM calculations.

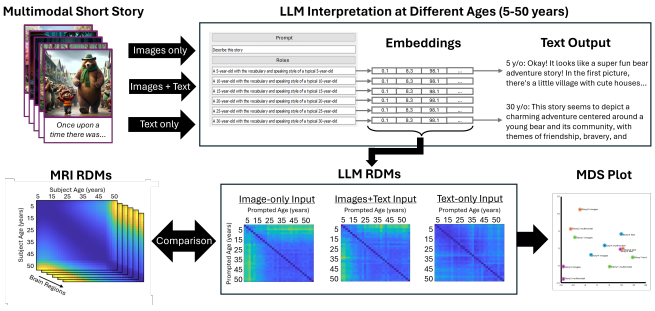

Figure 1: Processing pipeline summarizing the considered input modalities (image-only, image and text, text-only) and the LLM output analysis. Text output and embedding vectors were

Data Analysis. A second-degree squared Euclidean distance RDM was generated from the first degree LLM RDMs, and multi-dimensional scaling was applied to plot each input modality and each story in a 2D space. Following this analysis, to identify the brain regions with RDMs that best correlated with the RDMs obtained from LLM embeddings, a Bayesian optimization was performed across MRI RDMs for each given modality. We normalized each of the MRI RDMs by subtracting the global minimum value and dividing by the range of all RDM values. We then calculated the initial weights based on their Spearman correlation with the modality RDM. For each MRI RDM, the diagonal elements are set to zero, and the Spearman correlation between the lower left triangle of the MRI RDM and the lower left triangle of the modality RDM were computed. Negative correlations were replaced with zero. The weights were normalized to sum up to 1. The objective function calculates the Spearman correlation between the combined RDM (weighted sum of MRI RDMs) and the modality RDM.

The acquisition function used in the optimization is the Expected Improvement (EI) with $\xi = 0.01$ and $\kappa =$ None. The Gaussian process parameters are set with $\alpha = 10^{-5}$. The optimization is performed using the following equations:

$$\mathbf{w}^* =_{\mathbf{w}} \mathrm{EI}(\mathbf{w}) \tag{1}$$

$$\mathrm{EI}(\mathbf{w}) = E[\max(f(\mathbf{w}) - f^*, 0)] \tag{2}$$

where $\mathbf{w}$ is the vector of weights, $\mathbf{w}^*$ is the optimal set of weights, $\mathrm{EI}(\mathbf{w})$ is the Expected Improvement acquisition function, $f(\mathbf{w})$ is the objective function value for weights $\mathbf{w}$, and f^* is the current best objective function value. The number of iterations was set to 50 with 25 repetitions with different random seeds and we used the weights from the repetition with best performance to identify the optimal combination of brain areas.

Results

The LLM successfully associated character names presented in the text with the characters depicted in images, validating proper interpretation of multi-modal information. A validation prompt included images from one story with a text from a different story resulted in an output that incorporated elements from both stories. When considering the MDS shown in figure 2, image-only and image-text inputs produced more similar RDMs to each other for each story, whereas text-only input RDMs were clustered further. This MDS and the relative grouping also confirms the subjective findings through the semantic embeddings.

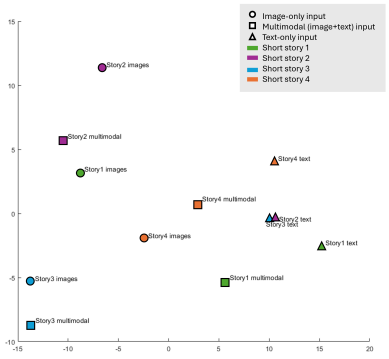

Figure 2: Multidimensional scaling of LLM RDMs for each short story and each input modality. The image-only and image+text RDMs are clustered more closely together than text-only RDMs.

Qualitatively, GPT 4 responses were modulated by age in prose, syntax, and detail of response. Spearman correlation revealed that image-only prompting correlated best with MRI changes (r=0.92), followed by image-text (r=0.85), then text-only (r=0.71) with p¡0.001 for all three cases. It's notable that several well-established brain networks emerge as strong correlates. For example, visual areas, lingual gyrus, temporal pole, pericalcarine regions, and regions associated with the default mode network showed good alignment with the prompts. More specifically, the cuneus and pre-cuneus regions and super parietal were activated across image-text, text-only, and image-only modalities. For the text-only modality, the temporal and occipital regions were highly correlated. Text-only had the highest spread of weights and in the gray matter volume (GMV), white matter volume (WMV), total cerebrum volume (TCV).

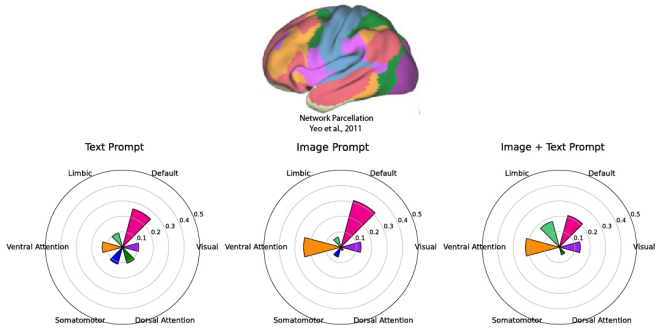

Figure 3: Rose plot depicting the optimal weighting scheme determined by Bayesian optimization for each input modality and the MRI RDM of each brain region organized into functional networks.

Discussion

In this study, we successfully implemented a framework for LLM interpretation of uni- and multi-modal data from the perspective of various human developmental stages. Multi-modal results demonstrated the LLM's versatility to synthesize information from multiple sources, integrating visual and textual information. The MDS scaling further underscored the text-only modality distinctiveness. This indicated its divergence from the image-only and image-text modalities. We also showed high (r¿0.9) correlation between changes in LLM activity and structural changes in the brain over developmental stages. Image-only input yielded the strongest correlation, which suggests that visual stimuli results in more similar LLM activity to human brain development compared to text-only stimuli. LLM interpretation of image-only input yielded high correlation with primary visual cortex (V1) structure. The Default Mode Network (DMN) is a network of brain regions that focus on an individual at rest with internally focused tasks such as memory retrieval (Hickok, 2009) and its development over age was shown to correlate with LLM activity for all three investigated input modalities. In the future we aim to explore the addition of auditory stimuli to expand the multi-modal component of the analysis. Additionally, we are developing a study in which human subjects at various ages will interpret the same stories that were passed to the LLM, allowing for a head-to-head comparison of human and LLM linguistics rather than comparing LLM behavior to structural imaging of the human brain.

Acknowledgments

Vanderbilt IDD-Reads Award, Vanderbilt SyBBURE Program, and Dr. Jonathan Ehrman for helpful feedback

References

Bethlehem, R. a. I., Seidlitz, J., White, S. R., Vogel, J. W., Anderson, K. M., Adamson, C., ... Alexander-Bloch, A. F. (n.d.). Brain charts for the human lifespan. , *604*(7906), 525–533. doi: 10.1038/s41586-022-04554-y

Caucheteux, C., & King, J.-R. (2022, February). Brains and algorithms partially converge in natural language processing. *Communications Biology*, *5*(1), 1–10. doi: 10.1038/s42003-022-03036-1

Hickok, G. (2009, September). The Functional Neuroanatomy of Language. *Physics of life reviews*, *6*(3), 121–143. doi: 10.1016/j.plrev.2009.06.001

Li, J., Karamolegkou, A., Kementchedjhieva, Y., Abdou, M., Lehmann, S., & Søgaard, A. (2023, October). *Structural Similarities Between Language Models and Neural Response Measurements* (No. arXiv:2306.01930). arXiv. doi: 10.48550/arXiv.2306.01930

McKenna, N., Li, T., Cheng, L., Hosseini, M. J., Johnson, M., & Steedman, M. (2023, October). *Sources of Hallucination by Large Language Models on Inference Tasks* (No. arXiv:2305.14552). arXiv.

Salewski, L., Alaniz, S., Rio-Torto, I., Schulz, E., & Akata, Z. (2023, November). *In-Context Impersonation Reveals Large Language Models' Strengths and Biases* (No. arXiv:2305.14930). arXiv. doi: 10.48550/arXiv.2305.14930