Human predictions of spoken language are more accurate and better aligned with neural activity than language model predictions

Thomas L. Botch (thomas.l.botch@dartmouth.edu)
Department of Psychological & Brain Sciences, Dartmouth College
3 Maynard Street, Hanover, NH 03755 USA

Emily S. Finn (emily.s.finn@dartmouth.edu)
Department of Psychological & Brain Sciences, Dartmouth College
3 Maynard Street, Hanover, NH 03755 USA

Abstract

Humans communicate through both spoken and written language, often switching between these modalities depending on their goals. We investigated the alignment of large language models (LLMs) and human participants (N=300) that predicted words within a story presented as either spoken language or written text. We found that LLM predictions were more similar to humans' predictions of written text, though humans' predictions of spoken language were the most accurate. By training encoding models to predict neural activity recorded with fMRI to the same auditory story, we showed that models based on human predictions of spoken language better aligned with observed brain activity compared to models based on either LLM predictions or human predictions of written text. These findings suggest that the structure of spoken language carries additional information relevant to human behavior and neural representations.

Keywords: Language; LLMs; Alignment; fMRI

Introduction

Large language models (LLMs) have provided researchers with tools to probe the functions that underpin efficient processing and representation of language (Linzen & Baroni, 2021). Many recent studies demonstrate that the mechanisms by which LLMs process and predict language relate to both human behavior (Wilcox, Gauthier, Hu, Qian, & Levy, 2020) and neural representations (Schrimpf et al., 2021; Goldstein et al., 2022). However, unlike LLMs, humans regularly switch between processing spoken and written language (Hulme & Snowling, 2014) and represent these structures through a common neural code (Deniz, Nunez-Elizalde, Huth, & Gallant, 2019).

While prior work has shown similarities between LLM next-word predictions and humans' next-word predictions for written text (Goldstein et al., 2022; Jacobs & McCarthy, 2020), no studies have yet investigated human predictions of upcoming words in an auditory stimulus. Given that spoken language carries extra-linguistic signals (i.e., prosody) used to infer speakers' intentions (Cole, 2015), and that these signals are integrated into neural representations (Khanna et al., 2024), it becomes particularly important to directly compare human behavior across modalities. We therefore aimed to identify differences in how humans perform next-word prediction in spoken versus written language, and whether these differences drive divergences from human neural activity and/or language models.

Here, we asked human participants to make next-word predictions during a real-world story presented as either spoken or written language. We leveraged these predictions to evaluate differences in humans' *behavioral alignment* to LLMs based on the modality of the stimulus. We then used these predictions to assess *representational alignment* — specifically, whether human predictions in either or both modalities were more closely aligned to neural activity than LLM predictions of the same story.

Materials and methods

Natural language fMRI dataset We analyzed fMRI data from 8 participants (3 female, age 21-34 years) who listened to 26 naturalistic stories while undergoing functional magnetic resonance imaging (fMRI). All participants listened to the same auditory stories (range: 7:10 min - 16:53 min) taken from The Moth podcast. One story (*wheretheressmoke*) was presented to participants across five separate scan sessions for the purpose of model evaluation.

Encoding models We trained voxel-wise encoding models (Huth, de Heer, Griffiths, Theunissen, & Gallant, 2016) to predict each participant's neural activity from features of the same natural language stimulus. We modeled each stimulus using four feature spaces: auditory (mel-frequency spectrogram), phoneme (CMU Pronouncing Dictionary), word-level semantic (word2vec), and contextualized semantic features (GPT2-XL). Encoding models were formalized as a banded-ridge regression (Nunez-Elizalde, Huth, & Gallant, 2019) that learns a separate regularization parameter for each feature space (including the separate transformer layers).

We trained each model using a leave-one-story-out cross-validation procedure (25 total folds). To evaluate the predictive performance of the trained models, we averaged neural responses across the five separate sessions of *wheretheressmoke* and correlated the predicted and true timeseries. We then identified significantly predicted voxels through a block-wise permutation test (10 TR blocks; n=1000 permutations; (LeBel, Jain, & Huth, 2021)).

Behavioral experiment We recruited two groups of human participants for the study (N=300 total). Both groups were presented with the validation story: *wheretheressmoke*. The first group of participants (spoken condition, N=150) listened to the story without seeing the transcript. The second group of participants (written condition, N=150) viewed the story word-by-word without hearing the audio track. At intervals spaced by a minimum of 10 words, participants were asked to generate a one-shot prediction for the upcoming word. Participants in both conditions provided responses to the same words, and the written words were presented at the spoken rate to mitigate timing differences between conditions. We focused our experiment on moments when LLMs either succeeded or failed at performing the same task (next-word prediction). To this end, we selectively sampled content words (e.g., removing stop-words, named-entities, etc.) based on the accuracy and entropy of GPT2-XL prediction distributions.

Alignment estimation We investigated the alignment of humans and LLMs performing next-word predictions. We defined behavioral alignment as the Kullback-Leibler (KL) divergence of human and LLM prediction distributions. To compare these distributions, we limited the LLM prediction distributions to the unique words predicted by human participants (in both

conditions). We also calculated the binary accuracy (exact match) of next-word predictions to the ground-truth word.

We then compared whether human- or LLM-predicted words provided better predictions of neural responses. To this end, we substituted predicted for ground-truth words within the original story to create three additional contextual feature spaces: humans' predictions from 1) spoken- and 2) written-language (both based on the most commonly predicted word in each modality), and 3) LLM predictions. We quantified representational alignment as the inverse mean squared error (MSE) of the predicted timeseries at the specific timepoints when a word was predicted. Within this definition, lower MSE indicates higher alignment of representations with neural activity. We then contrasted the MSE of these predicted timeseries to determine which of the three representational spaces better fit brain responses.

Results

Across both stimulus modalities (spoken or written), human predictions were more accurate than LLM predictions (Fig. 1A; spoken-model: $p < 0.001$; written-model: $p = 0.08$; spoken-written: $p = 0.29$). We then evaluated whether LLM prediction distributions well represented the distributions of words predicted by human participants. On average, the LLM distribution exhibited significantly lower KL divergence when evaluated against the written-language distribution as compared to the spoken-language distribution (Fig. 1B; $t(472) = 2.39, p = 0.017$). This suggests that LLM prediction patterns were more similar to humans in the written modality as compared to the spoken modality.

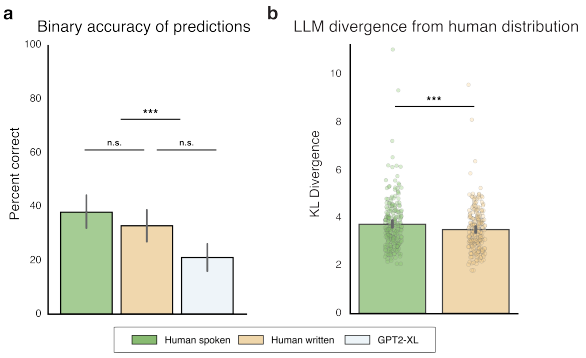

Figure 1: (a) Accuracy of next-word predictions. (b) LLM prediction distributions better fit distributions of humans predicting written text. *** $p < 0.001$, n.s. $p > 0.05$.

To understand if these behavioral differences are recapitulated in the alignment of neural representations, we examined the accuracy of human- or LLM-predicted words on predicting brain activity. Across the majority of significantly predicted voxels, we found that human predictions in both stimulus modalities better aligned with human neural representations than LLM predicted words (Fig. 2A; $q_{FDR} < 0.05$).

We then compared the representational alignment between

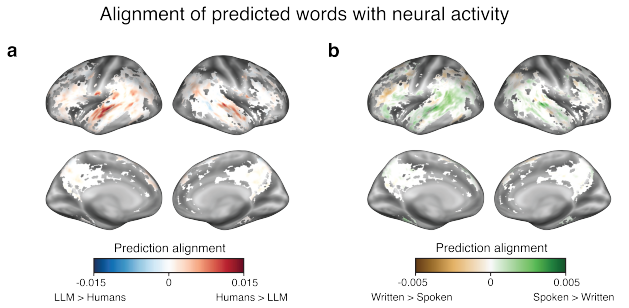

Figure 2: Difference in prediction alignment between (a) humans and model predictions and (b) spoken and written language predictions. All plots are thresholded by encoding model significance at each voxel ($q_{FDR} < 0.05$).

human predictions of spoken and written language. We found that words from human predictions of spoken language broadly exhibited greater alignment to brain activity than predictions of written language (Fig. 2B; $q_{FDR} < 0.05$). Interestingly, spoken language predictions demonstrated greater alignment than written language across the majority of auditory and language related regions. This result provides a parallel to the divergence observed in human behavior and suggests that human predictions of spoken language are more representative of neural responses, at least during auditory perception.

In sum, human predictions of both spoken and written language were more accurate than LLMs. However, humans predictions of written text showed greater alignment with LLMs than predictions of spoken language. These differences in behavioral alignment were recapitulated in the alignment of predicted words with human neural representations. Human predictions better aligned with brain activity than LLM predictions, and predictions of spoken language were more aligned with neural activity than those of written language. Together, these findings suggest that LLM predictions are less aligned with both human behavior and neural activity than previously assumed and highlight how the rich, multimodal nature of spoken language may aid situational understanding to enable more accurate predictions.

References

Cole, J. (2015, February). Prosody in context: a review. *Language, Cognition and Neuroscience*, *30*(1-2), 1–31. doi: 10.1080/23273798.2014.963130

Deniz, F., Nunez-Elizalde, A. O., Huth, A. G., & Gallant, J. L. (2019, September). The Representation of Semantic Information Across Human Cerebral Cortex During Listening Versus Reading Is Invariant to Stimulus Modality. *The Journal of Neuroscience*, *39*(39), 7722–7736. doi: 10.1523/JNEUROSCI.0675-19.2019

Goldstein, A., Zada, Z., Buchnik, E., Schain, M., Price, A., Aubrey, B., ... Hasson, U. (2022, March). Shared computational principles for language processing in humans

and deep language models. *Nature Neuroscience*, *25*(3), 369–380. doi: 10.1038/s41593-022-01026-4

Hulme, C., & Snowling, M. J. (2014, January). The interface between spoken and written language: developmental disorders. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1634), 20120395. doi: 10.1098/rstb.2012.0395

Huth, A. G., de Heer, W. A., Griffiths, T. L., Theunissen, F. E., & Gallant, J. L. (2016, April). Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, *532*(7600), 453–458. doi: 10.1038/nature17637

Jacobs, C. L., & McCarthy, A. D. (2020). The human unlikeness of neural language models in next-word prediction. In *Proceedings of the The Fourth Widening Natural Language Processing Workshop* (pp. 115–115). Seattle, USA: Association for Computational Linguistics. doi: 10.18653/v1/2020.winlp-1.29

Khanna, A. R., Muñoz, W., Kim, Y. J., Kfir, Y., Paulk, A. C., Jamali, M., ... Williams, Z. M. (2024, January). Single-neuronal elements of speech production in humans. *Nature*. doi: 10.1038/s41586-023-06982-w

LeBel, A., Jain, S., & Huth, A. G. (2021, December). Voxelwise Encoding Models Show That Cerebellar Language Representations Are Highly Conceptual. *The Journal of Neuroscience*, *41*(50), 10341–10355. doi: 10.1523/JNEUROSCI.0118-21.2021

Linzen, T., & Baroni, M. (2021, January). Syntactic Structure from Deep Learning. *Annual Review of Linguistics*, *7*(1), 195–212. (arXiv:2004.10827 [cs]) doi: 10.1146/annurev-linguistics-032020-051035

Nunez-Elizalde, A. O., Huth, A. G., & Gallant, J. L. (2019, August). Voxelwise encoding models with non-spherical multivariate normal priors. *NeuroImage*, *197*, 482–492. doi: 10.1016/j.neuroimage.2019.04.012

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

Wilcox, E. G., Gauthier, J., Hu, J., Qian, P., & Levy, R. (2020). On the Predictive Power of Neural Language Models for Human Real-Time Comprehension Behavior. (Publisher: arXiv Version Number: 1) doi: 10.48550/ARXIV.2006.01912