Inducing bias from bilingual brains into language models

Anuja Negi (anuja.negi@tu-berlin.de)[1, 2], Subba Reddy Oota[1], Fatma Deniz[1, 2]
[1]Technical University of Berlin, Germany [2] Bernstein Center for Computational Neuroscience Berlin, Germany

Abstract

Recent studies have shown that inducing bias from neural data in language models can enhance their ability to encode brain activity and improve performance on language tasks. However, these approaches have mainly focused on a single language. Given recent evidence that semantic representations are shared across languages in the bilingual brain, we ask whether brain-informed fine-tuning can reveal latent multilingual capabilities in language models. To test this, we fine-tune pretrained monolingual Transformer models (English and Chinese BERT) using fMRI data from bilingual individuals. We find that fine-tuning improves downstream performance not only in the language used for training but also in the other language, indicating cross-linguistic generalization. Further, the encoding performance of the fine-tuned model across other participants remains the same, suggesting that the brain bias introduced by fine-tuning is shared across bilingual individuals.

Keywords: fMRI, Language Models, Cognitive Neuroscience, Brain Alignment, Encoding Models, Multilingualism

Introduction

Recent research has demonstrated that text-based language models can predict human brain activity during language processing, suggesting parallels between artificial and neural language representations (Wehbe et al., 2014; Jain & Huth, 2018; Toneva & Wehbe, 2019; Schrimpf et al., 2021; Caucheteux & King, 2022; Goldstein et al., 2022; Karamolegkou et al., 2023). Although these models effectively capture patterns in brain activity, they are not originally trained to capture language representations in the human brain.

Fine-tuning text-based language models with brain data has been shown to improve their neural encoding performance without negatively affecting downstream natural language processing (NLP) tasks (Schwartz et al., 2019). Similarly, brain-informed fine-tuning approaches have recently been applied to speech-based language models, yielding improvements in downstream NLP tasks (Moussa et al., 2024; Vattikonda et al., 2025). However, prior research has primarily focused on a single language (usually English), overlooking the prevalence of bilingualism in human populations. This limitation is particularly notable given recent neuroscientific evidence revealing that bilingual individuals have shared semantic representations across languages (Chen et al., 2024).

Our study investigates whether fine-tuning language models with brain data from bilingual individuals can reveal latent multilingual capabilities. Specifically, we fine-tune monolingual text-based language models using bilingual fMRI (functional magnetic resonance imaging) data collected during naturalistic language. We find that fine-tuning language models with bilingual brain data improves language models' downstream NLP performance not only in the language used for fine-tuning but also in the second language. Furthermore, this fine-tuning does not degrade their brain encoding performance. Our results contribute to the alignment between brain and computational multilingual language representations, offering insights into the development of brain-inspired multilingual NLP systems.

Methodology

Dataset We used functional magnetic resonance imaging (fMRI) to record human brain responses from six bilingual participants (three males and three females) fluent in both Mandarin Chinese (native) and English (non-native). Each participant read 11 spoken stories (narratives) from The Moth Radio Hour word by word, in English (en) and Chinese (zh) (previously used by (Chen et al., 2024)) in separate sessions. Of these narratives, seven were used for fine-tuning the models, three were utilized for training the voxelwise encoding models, and one was reserved for testing. The same set of narratives was used consistently across both the English and Chinese conditions. The BOLD responses were z-scored before usage.

Text-based Language Models We fine-tune two monolingual Transformer-based text-based language models: English BERT (bert-en) and Chinese-BERT (bert-zh) (Devlin et al., 2019). Both models share an identical architecture comprising 12 Transformer layers with a hidden size of 768, differing only in their pretraining language datasets. Pretrained model checkpoints were obtained from HuggingFace (Wolf et al., 2020).

Fine-tuning Language models with Brain Data We performed supervised full fine-tuning of pretrained language models by updating all model weights using fMRI BOLD (blood-oxygenation-level–dependent imaging) responses as targets. Transcripts of the training narratives were provided as inputs to the model, using a sequence length of 20 tokens. For each input word, we extracted the representation corresponding to the last token from the last hidden layer. These representations were then passed through a pooling layer designed to perform downsampling and temporal delaying. The pooled representations were projected into voxel space through a linear layer, thereby predicting per-voxel BOLD responses. Training was performed using batched inputs and optimized with AdamW (Loshchilov & Hutter, 2017) (lr = 1e-4) for 30 epochs. Our training objective minimized the NT-Xent (Normalized Temperature-Scaled Cross-Entropy) loss (Sohn, 2016) between the predicted and actual BOLD responses.

	GLUE tasks (english)										CLUE tasks (chinese)						
	CoLA	SST-2	MRPC (Acc.)	STS-B (Pear.)	QQP (Acc.)	MNLI-m	MNLI-mm	QNLI	RTE	WNLI	afqmc	cmnli	csl	iflytek	tnews	chid	c3
vanilla-en	53.38	92.08	79.41	88.06	90.84	84.38	84.64	91.45	67.15	49.30	69.00	68.34	71.20	47.86	50.92	10.66	42.64
ft-en	**55.75**	**93.12**	**80.88**	**88.10**	**90.91**	**84.69**	**84.71**	**91.65**	66.43	**50.07**	**69.00**	**68.63**	**71.43**	46.79	50.60	**10.66**	41.12
vanilla-zh	42.17	85.09	78.43	80.06	86.40	73.42	74.33	82.56	61.01	56.34	75.25	80.50	80.18	60.25	56.44	10.66	49.74
ft-zh	**43.25**	**86.12**	78.19	79.44	**86.48**	**73.65**	**74.82**	**82.99**	58.48	**56.34**	**75.25**	**80.83**	**80.73**	**60.29**	56.28	**10.66**	**49.92**

Table 1: Performance on downstream NLP tasks. Bolded values indicate equal or improved performance compared to the corresponding pretrained (non-fine-tuned) model.

Models were fine-tuned separately for each participant and language.

Voxelwise encoding model fitting We used a voxelwise encoding modeling (VM) approach to evaluate whether language model representations better predict brain responses before or after fine-tuning. We extract contextual embeddings from the ninth hidden layer of the model. Ridge regression was used to determine how the embedding is represented in each voxel (Wu et al., 2006; Naselaris et al., 2011). Prediction accuracy was quantified by calculating the Pearson correlation coefficient (r) between predicted and recorded BOLD responses on the held-out test narrative. A separate VM was fit for each voxel, participant, and language.

Downstream NLP tasks To evaluate the effects of fine-tuning on language model behavior, we assess performance on standard NLP benchmarks. For English, we use the GLUE benchmark (Wang et al., 2018), and for Chinese, we use the CLUE benchmark (Xu et al., 2020).

Results

Improved Downstream Performance in Fine-Tuned Language Fine-tuning language models with bilingual brain data improves most downstream task performance (compared to its vanilla counterpart) in the same language used for fine-tuning. As shown in Table 1, 9/10 GLUE and 6/7 CLUE tasks improve. These results suggest that brain-informed changes to language representations can strengthen a model's language processing capabilities in the fine-tuned language.

Cross-Linguistic Generalization of Brain-Induced Bias Downstream performance improvements extend beyond the language used during fine-tuning. We observe that models fine-tuned in one language (e.g., Chinese) also perform better on downstream tasks in the other language (e.g., English). Table 1 shows improvements in 7/10 GLUE tasks when weights from the Bert-ft-zh model are transferred to Bert-en, and in 4/7 CLUE tasks for the reverse transfer. These findings suggest that the bias introduced by fine-tuning with brain data captures language-general semantic structure that facilitates effective cross-linguistic transfer.

Effects on Brain Encoding Performance and Generalization We evaluated VM performance before and after fine-tuning to assess changes in brain alignment and to test for possible overfitting to individual participants. Figure 1 shows cortical flatmaps of changes in encoding performance with Bert-ft-en for (a) the fine-tuned subject and (b) a different subject. In both cases, no degradation in encoding performance is observed across the cortex. This is consistently observed across all participants and for the Bert-zh model. These findings suggest that the changes introduced by fine-tuning are not participant-specific but reflect some shared representations across bilingual individuals.

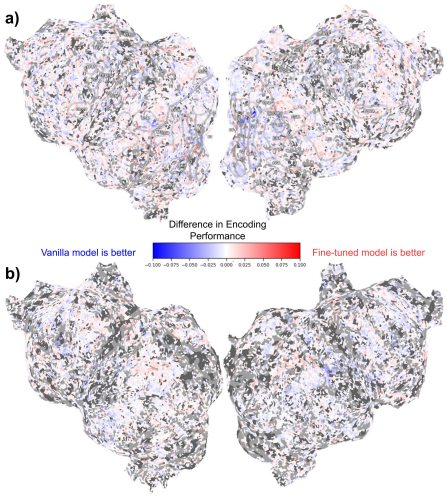

Figure 1: Change in encoding performance after fine-tuning for (a) the same subject and (b) a different subject.

Discussion and Conclusion

In this study, we investigated whether language models exhibit latent multilingual capabilities when fine-tuned with bilingual brain data. We fine-tuned monolingual language models using fMRI recordings from bilingual participants during naturalistic language comprehension and evaluated the fine-tuned models on downstream tasks and brain encoding performance.

Our results show that brain-informed fine-tuning improves performance on NLP tasks in the language used for fine-tuning. Notably, monolingual models exhibited enhanced performance in the non-fine-tuned language, suggesting that brain-informed fine-tuning induces generalizable semantic structure not tied to a specific language. Further, our fine-tuning approach does not affect the model's ability to encode brain responses across participants. These findings suggest that monolingual models possess a latent capacity for multilingual processing that can be revealed through brain-informed fine-tuning. In future work, we aim to test this method across a wider range of language model architectures and extend this approach to multilingual models.

Acknowledgement

We thank Lily Gong for preprocessing the fMRI data, and Anwar Nunez-Elizalde and Mathis Lamarre for valuable discussions. This work was funded by grants from the German Federal Ministry of Education and Research (BMBF; Grant no. 01GQ1906) and the European Research Council (ERC; Grant no. 101042567).

References

Caucheteux, C., & King, J.-R. (2022). Brains and algorithms partially converge in natural language processing. *Communications biology*, *5*(1), 134.

Chen, C., Gong, X. L., Tseng, C., Klein, D. L., Gallant, J. L., & Deniz, F. (2024). Bilingual language processing relies on shared semantic representations that are modulated by each language. *bioRxiv*, 2024–06.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the north american chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)* (pp. 4171–4186).

Goldstein, A., Zada, Z., Buchnik, E., Schain, M., Price, A., Aubrey, B., ... others (2022). Shared computational principles for language processing in humans and deep language models. *Nature neuroscience*, *25*(3), 369–380.

Jain, S., & Huth, A. (2018). Incorporating context into language encoding models for fmri. *Advances in neural information processing systems*, *31*.

Karamolegkou, A., Abdou, M., & Søgaard, A. (2023). Mapping brains with language models: A survey. *arXiv preprint arXiv:2306.05126*.

Loshchilov, I., & Hutter, F. (2017). Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Moussa, O., Klakow, D., & Toneva, M. (2024). Improving semantic understanding in speech language models via brain-tuning. *arXiv preprint arXiv:2410.09230*.

Naselaris, T., Kay, K. N., Nishimoto, S., & Gallant, J. L. (2011). Encoding and decoding in fmri. *Neuroimage*, *56*(2), 400–410.

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

Schwartz, D., Toneva, M., & Wehbe, L. (2019). Inducing brain-relevant bias in natural language processing models. *Advances in neural information processing systems*, *32*.

Sohn, K. (2016). Improved deep metric learning with multi-class n-pair loss objective. *Advances in neural information processing systems*, *29*.

Toneva, M., & Wehbe, L. (2019). Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain). *Advances in neural information processing systems*, *32*.

Vattikonda, N., Vaidya, A. R., Antonello, R. J., & Huth, A. G. (2025). Brainwavlm: Fine-tuning speech representations with brain responses to language. *arXiv preprint arXiv:2502.08866*.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. R. (2018). Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

Wehbe, L., Vaswani, A., Knight, K., & Mitchell, T. (2014). Aligning context-based statistical models of language with brain activity during reading. In *Proceedings of the 2014 conference on empirical methods in natural language processing (emnlp)* (pp. 233–243).

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., ... others (2020). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations* (pp. 38–45).

Wu, M. C.-K., David, S. V., & Gallant, J. L. (2006). Complete functional characterization of sensory neurons by system identification. *Annu. Rev. Neurosci.*, *29*(1), 477–505.

Xu, L., Hu, H., Zhang, X., Li, L., Cao, C., Li, Y., ... others (2020). Clue: A chinese language understanding evaluation benchmark. *arXiv preprint arXiv:2004.05986*.