Brain-Informed Fine-Tuning for Improved Multilingual Understanding in Language Models

Anuja Negi [1, 2] ,*Subba Reddy Oota[1],*Manish Gupta [3], Fatma Deniz[1, 2]
[1]Technical University of Berlin, Germany,
[2] Bernstein Center for Computational Neuroscience Berlin, Germany,
[3] Microsoft

Abstract

Recent studies have demonstrated that fine-tuning language models with brain data can improve their semantic understanding, although these findings have so far been limited to English. Interestingly, similar to the shared multilingual embedding space of pretrained multilingual language models, human studies provide strong evidence for a shared semantic system in bilingual individuals. Here, we investigate whether fine-tuning language models with bilingual brain data changes model representations in a way that improves them across multiple languages. To test this, we fine-tune monolingual and multilingual language models using brain activity recorded while bilingual participants read stories in English and Chinese. We then evaluate how well these representations generalize to the bilingual participants' first language, their second language, and several other languages that the participant is not fluent in. We assess the fine-tuned language models on brain encoding performance and downstream NLP tasks. Our results show that bilingual brain-informed fine-tuned language models outperform their vanilla (pretrained) counterparts in both brain encoding performance and most downstream NLP tasks across multiple languages. These findings suggest that brain-informed fine-tuning improves multilingual understanding in language models, offering a bridge between cognitive neuroscience and NLP research.

1 Introduction

Recent research has demonstrated that representations extracted from text-based Transformer language models can be used to accurately predict human brain activity evoked during language processing, suggesting parallels between artificial and brain language representations (Wehbe et al., 2014b; Jain & Huth, 2018; Toneva & Wehbe, 2019; Schrimpf et al., 2021; Caucheteux & King, 2022; Goldstein et al., 2022; Karamolegkou et al., 2023). Although these models accurately predict patterns of brain activity, they have not been originally pretrained to do so. Several previous studies hypothesized that fine-tuning language models with brain data can lead to representations that are better aligned with brain activity (Schwartz et al., 2019; Moussa et al., 2025; Vattikonda et al., 2025). However, these efforts have largely focused on monolingual language models and monolingual brain data, usually trained and evaluated only in English. This overlooks the widespread prevalence of bilingualism and multilingualism in the human population (Grosjean, 2024). The limitation is particularly significant given recent neuroscientific evidence revealing that bilingual individuals have shared semantic representations across languages (Chen et al., 2024b; Francis, 2005), hinting at a shared component for the processing of different languages in the human brain (de Varda et al., 2025). Complementary NLP research has shown that multilingual language models operate in a shared, language-agnostic conceptual space (Wendler et al., 2024; Schut et al., 2025). This raises the question of *whether fine-tuning language models with bilingual brain data can elicit multilingual capabilities in language models.*

*equal contributors

Preprint. Under review.

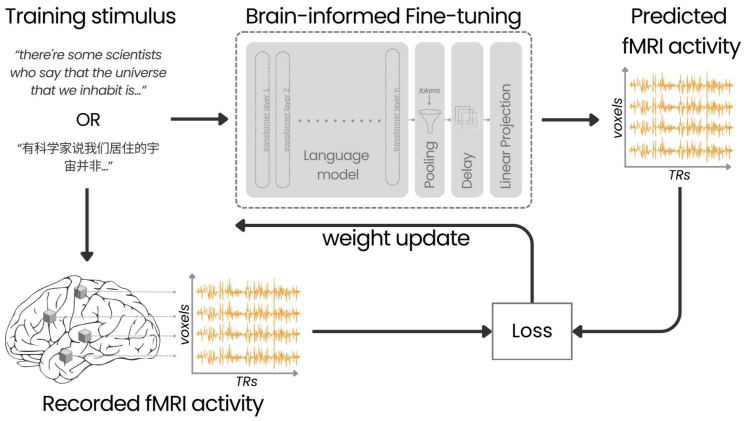

Figure 1: **Brain-informed fine-tuning pipeline**. Participants read naturalistic stories in English or Chinese while fMRI responses were recorded. The corresponding transcripts were fed into a pretrained language model. Representations from the final token of the last hidden layer were downsampled and temporally delayed to align with the fMRI acquisition. These were then projected to voxel space via a linear layer to predict brain activity. The loss between predicted and recorded responses was backpropagated through all layers for full model fine-tuning.

In this study, we use brain recordings of bilingual participants reading the same naturalistic stories in English and Chinese and investigate whether fine-tuning language models with brain data from bilingual participants can elicit multilingual capabilities in language models. For this, we first introduce a novel, full brain-informed fine-tuning pipeline (as shown in Fig. 1), which can be trained using data from the whole-brain, or from language- or semantically-selective brain regions. Second, we use brain recordings of bilingual participants to fine-tune two monolingual language models (BERT for English and Chinese (Devlin et al., 2019)) and four multilingual language models (mBERT (Devlin et al., 2019), XLM-R (Conneau et al., 2020), XGLM (Lin et al., 2022), LLaMA-3.2 (Touvron et al., 2023)). Third, we investigate how monolingual and multilingual language models change when fine-tuned with bilingual brain data. We evaluate this along two axes: brain encoding and downstream NLP tasks performance. To assess generalization across languages (zero-shot cross-lingual transfer), we evaluate the brain-informed fine-tuned language models on downstream NLP benchmarks not only in the models' fine-tuned language but also in the bilingual participants' other language (cross-language transfer between known languages) and several other languages that the participant is not fluent in (and also not used in fine-tuning; referred to as 'unseen languages'), e.g., German and French. Fine-tuning is done separately for each participant, and we assess encoding performance both within the same participant and across the other participants.

Brain-informed fine-tuning with bilingual brain data reveals several key conclusions: (1) Brain encoding performance improves after fine-tuning, even when evaluated on participants not used in training. This suggests that the brain bias introduced is not specific to a participant but reflects some shared bilingual representations. (2) Both monolingual and multilingual language models show improved performance on downstream NLP tasks in the fine-tuned language, the participant's other language, and in several other unseen languages, indicating that brain-informed fine-tuning elicits generalizable semantic structure not tied to any one language. (3) To ascertain whether our results are specific to fine-tuning with bilingual brain data, and not a general outcome of fine-tuning language models with brain data, we performed the same analyses using brain data of a monolingual individual. Results suggest that the observed effects are indeed driven by bilingual brain representations. These findings suggest that bilingual brain-informed fine-tuning improves multilingual understanding in text-based language models. Our results contribute to the alignment between brain and artificial multilingual language representations, offering insights into the development of brain-inspired multilingual NLP systems.

We make the following contributions: (1) To the best of our knowledge, this is the first study to perform brain-informed fine-tuning using bilingual brain data, applying it to both monolingual and multilingual language models. (2) We introduce a novel brain-informed fine-tuning pipeline in which downsampling and temporal-delay modeling are integrated within the pipeline itself. This contrasts with previous brain-based fine-tuning studies, where these were modeled as preprocessing steps before fine-tuning the model. (3) We evaluate the performance of brain-informed fine-tuned monolingual

and multilingual language models on downstream NLP tasks in both English and Chinese. We will release the code upon publication of this paper.

2 Related Work

Fine-tuning of language models with naturalistic brain data. Our work builds on the brain-tuning approach introduced by Schwartz et al. (2019); Moussa et al. (2025); Vattikonda et al. (2025), which fine-tunes pretrained Transformer-based language models using brain data to integrate brain-relevant information. Schwartz et al. (2019) demonstrated improved brain encoding and NLP task performance using brain data from monolingual English readers, while Moussa et al. (2025); Vattikonda et al. (2025) extended this to speech-based models to enhance semantic representations. Our study complements these by exploring bilingual brain-informed fine-tuning and analyzing how monolingual and multilingual models change when trained with bilingual brain data.

Multilingual language models and brain alignment. Our work aligns with a growing body of research examining the alignment between human brain activity and language models. Several studies have shown that text-based models can predict brain responses to written and spoken stimuli with high accuracy (Wehbe et al., 2014a; Jain & Huth, 2018; Toneva & Wehbe, 2019; Deniz et al., 2019; Abdou et al., 2021; Toneva et al., 2022; Antonello et al., 2021; Oota et al., 2022; Aw & Toneva, 2023; Oota et al., 2024b; Lamarre et al., 2022; Chen et al., 2024a; Oota et al., 2024a). Recent efforts have extended this to multilingual Transformer-based models using brain data from tasks in multiple languages (de Varda et al., 2025), though most studies remain monolingual in design, with the exception of Chen et al. (2024a), who explored bilingual processing using similar English and Chinese stimuli. Our study supplements this by investigating bilingual brain alignment through brain-informed fine-tuning and its impact on downstream NLP tasks in both languages. Detailed related work is in Appendix A.

3 Methodology

3.1 Naturalistic Brain Imaging Dataset

Bilingual fMRI dataset. Blood oxygen level-dependent (BOLD) responses were collected using fMRI from six bilingual participants (three males, three females), all native speakers of Mandarin Chinese and fluent in English as a second language. Participants read 11 narrative stories from The Moth Radio Hour (Huth et al., 2016), presented word-by-word in separate scanning sessions for English (en) and Chinese (zh). Each story is 10-15 minutes long. Of the 11 stories, seven were used for fine-tuning the language models (covering 2756 TRs), three were used for training voxelwise encoding models (1117 TRs), and one story was reserved for testing (291 TRs). The same set of stories are used in each language to ensure comparability across languages.

Monolingual fMRI dataset. To test whether the effects of brain-informed fine-tuning are specific to bilingual brains, we also use fMRI data from one English-monolingual participant (male). The participant was presented with the same English narrative stories under the same experimental design and scanning parameters. This allows direct comparison between models fine-tuned with bilingual and monolingual brain data. If similar improvements are seen with monolingual brain fine-tuning, it suggests that the effects are driven by general brain representations rather than shared semantic representations in bilinguals. More details about datasets and preprocessing are in Appendix B.

3.2 Text-based Language Models

We fine-tuned monolingual and multilingual Transformer-based text-based models to investigate the effects of brain-informed fine-tuning. All pretrained model checkpoints were obtained from HuggingFace (Wolf et al., 2020).

Monolingual pretrained language models. We fine-tuned two monolingual models: English BERT (BERT-en) and Chinese BERT (BERT-zh) (Devlin et al., 2019). Both models share the same architecture, comprising 12 Transformer layers with a hidden dimension size of 768, differing only in the language of their pretraining corpora.

Multilingual pretrained language models. We fine-tuned four multilingual Transformer-based models: mBERT (Devlin et al., 2019), XLM-R (Conneau et al., 2020), XGLM (Lin et al., 2022), and LLaMA-3.2 (Touvron et al., 2023). These models represent three distinct architecture types: encoder-based (mBERT, XLM-R), cross-lingual pretrained (XGLM), and decoder-based (LLaMA-3.2). Representations were extracted from the base versions of mBERT, XLM-R, and XGLM, and from the 1B parameter version of LLaMA-3.2. mBERT, XLM-R, and XGLM consist of 12 layers with a hidden dimension size of 768, while LLaMA-3.2 has 16 layers with a hidden dimension size of 2048. Results and analyses using XLM-R, XGLM and LLaMA-3.2 are reported in the Supplementary.

3.3 Brain-informed Fine-Tuning Approach

We performed supervised full fine-tuning of pretrained language models using fMRI BOLD responses as targets. Models were fine-tuned separately for each participant and language. An overview of the fine-tuning pipeline is shown in Fig. 1. We fine-tune language models using fMRI responses from the whole-brain, language-selective (from (Fedorenko et al., 2010)), or semantically-selective brain regions. Details on fine-tuning with language-selective and semantically-selective brain regions are described in Appendix C.

Model architecture. Text transcripts of the narrative stimuli were provided as inputs to the pretrained language model, using a sequence length of 20 tokens. We then extract the representations of the last token from the last hidden layer for each word. This representation is given as input to a dropout layer. This is followed by differentiable 3-lobe Lanczos (Huth et al., 2016) interpolation to downsample to the fMRI sampling rate. To model the delayed hemodynamic response, we apply finite impulse response (FIR) filtering by concatenating representations delayed by 2, 4, 6, and 8 seconds. The resulting representations are then passed through a linear projection layer to predict voxelwise BOLD responses. Previous brain-based fine-tuning studies (Moussa et al., 2025; Vattikonda et al., 2025) typically treated downsampling and temporal-delay modeling as separate preprocessing steps before model fine-tuning. In contrast, our method integrates these components directly into the model architecture, enabling end-to-end optimization.

Training protocol. We fine-tuned the models using the AdamW optimizer (Loshchilov & Hutter, 2017) with a learning rate of 1e-4 and weight decay of 1e-3 for 30 epochs with a batch size of 32. A ReduceLROnPlateau scheduler was used to adjust the learning rate based on validation loss. We used mixed-precision training for computational efficiency and applied early stopping with a patience of 5 epochs based on validation performance. Implementation details are in Appendix G.

Training objective. Our training objective minimized the NT-Xent (Normalized Temperature-Scaled Cross-entropy) loss (Sohn, 2016) between the predicted and actual BOLD responses. We also experimented with alternative loss functions such as MSE, ridge loss, spatial loss, and a combination of all (referred to as hybrid), and found NT-Xent to perform best across models and experiments. Results from other losses are reported in Supplementary. We fine-tuned all layers of the language models and propagated the loss backward to update both the projection and transformer layers.

In addition to comparing our brain-informed fine-tuned models to their pretrained counterparts in both monolingual and multilingual settings, we include baselines for comprehensive comparison: TR-shuffled fMRI, multilingual model representations, and monolingual brain data as fine-tuning targets. This experiment was done to clarify the impact of bilingual brain data on the resulting fine-tuned model representations. Details are reported in Appendix C.3.

3.4 Voxelwise Encoding Modeling

To evaluate whether brain-informed fine-tuning alters the alignment between language model representations and brain activity, we assess voxelwise encoding performance before and after fine-tuning. Voxelwise encoding models (VEM) estimate, for each voxel, a linear mapping from an embedding to the observed fMRI responses (Huth et al., 2016; Deniz et al., 2019). If fine-tuning with bilingual brain data improves encoding performance, we interpret this as evidence that the language model's internal representations have become more brain-like, indicating successful alignment. If encoding performance remains unchanged, this suggests that fine-tuning preserves the existing brain alignment of the representations without introducing degradation. A decrease in performance would imply that fine-tuning introduces distortions that reduce the language model's ability to predict brain responses, thus reducing its alignment with the brain.

We used representations extracted from layer 7 (found to yield the highest encoding performance based on validation data) as the embedding. Embeddings and brain responses were z-scored across time separately for each of the three stories not used for fine-tuning. Embeddings were downsampled to match the fMRI acquisition rate using 3-lobe Lanczos interpolation. Next, to account for the delayed hemodynamic response, each embedding was passed through a finite impulse response (FIR) filter with four delays. Specifically, delayed copies of each feature at 1, 2, 3, and 4 TRs (2, 4, 6, and 8 seconds) were concatenated. Banded ridge regression (Nunez-Elizalde et al., 2019) was used to determine how each embedding is represented in each voxel (Wu et al., 2006; Naselaris et al., 2011). We performed 5-fold cross-validation to find optimal regularization parameters. Encoding performance was quantified by calculating the Pearson correlation coefficient (r) between predicted and recorded BOLD responses. The held-out story was used to estimate the encoding performance. A separate VEM was fit for each voxel, participant, and language model variant (before and after brain-informed fine-tuning). No sample size calculations were performed, as each subject serves as a full replication of the results.

Cross-participant transfer of encoding performance. We evaluated whether brain-informed fine-tuned model representations generalize across participants. We fine-tuned the language model using brain data from a one bilingual participant and evaluated its encoding performance using VEMs on others. Improved transfer would suggest that changes introduced by brain-informed fine-tuning are not participant-specific but reflect shared representations across bilingual individuals.

3.5 Downstream NLP tasks

To evaluate how brain-informed fine-tuning influences language model behavior, we assessed its performance on standard NLP benchmarks before and after fine-tuning. For English, we used 9 tasks from GLUE benchmark (Wang et al., 2018), and for Chinese, we used 7 tasks from CLUE benchmark (Xu et al., 2020). Both monolingual and multilingual models were evaluated on these benchmarks. To investigate cross-linguistic generalization to unseen languages, we additionally evaluated multilingual models on 3 tasks from XGLUE benchmark (Liang et al., 2020) and 3 tasks from XTREME benchmark (Hu et al., 2020). Details about these benchmarks are in Appendix D.

Fine-tuning and evaluation in the same language. To assess how brain-informed fine-tuning of language models affects performance on NLP tasks within the same language, we fine-tuned language models using fMRI data in English or Chinese. We then evaluated these language models on NLP benchmarks in the corresponding language. English and Chinese brain data are used to fine-tune monolingual (BERT-en, BERT-zh) and multilingual (mBERT) language models, which are then evaluated on GLUE and CLUE, respectively. Any performance increase observed in this setting would imply that brain-informed fine-tuning not only changes language model representations to better align with the brain, but that these modified representations are better for NLP tasks within the trained language.

Cross-language transfer between known languages. To test whether brain-informed fine-tuning transfers to a bilingual participant's other language, we evaluate models on the participant's other language after fine-tuning on the first. For example, weights from BERT-en fine-tuned with English brain data are transferred to BERT-zh and evaluated on Chinese tasks, and vice versa. Improved performance in this setting would suggest that brain-informed fine-tuning introduces the bilingual brain's shared semantic representations into the model, enabling cross-language transfer.

Zero-shot transfer to unseen languages. To evaluate the broader generalizability of brain-informed fine-tuned models, we tested multilingual models on five additional languages (German, French, Spanish, Japanese, and Korean) that were neither used for fine-tuning nor known to the participant. Specifically, we fine-tuned mBERT on English brain data and evaluated its performance on the XGLUE (Liang et al., 2020) and XTREME (Hu et al., 2020) benchmarks. This setup allows us to assess whether brain-informed fine-tuning can introduce language-agnostic representations that support broader cross-linguistic transfer.

4 Results

4.1 Effects on Brain Encoding Performance and Generalization

We tested whether brain-informed fine-tuning improves brain encoding performance and if these possible improvements are generalizable across participants. We compared the VEM performance

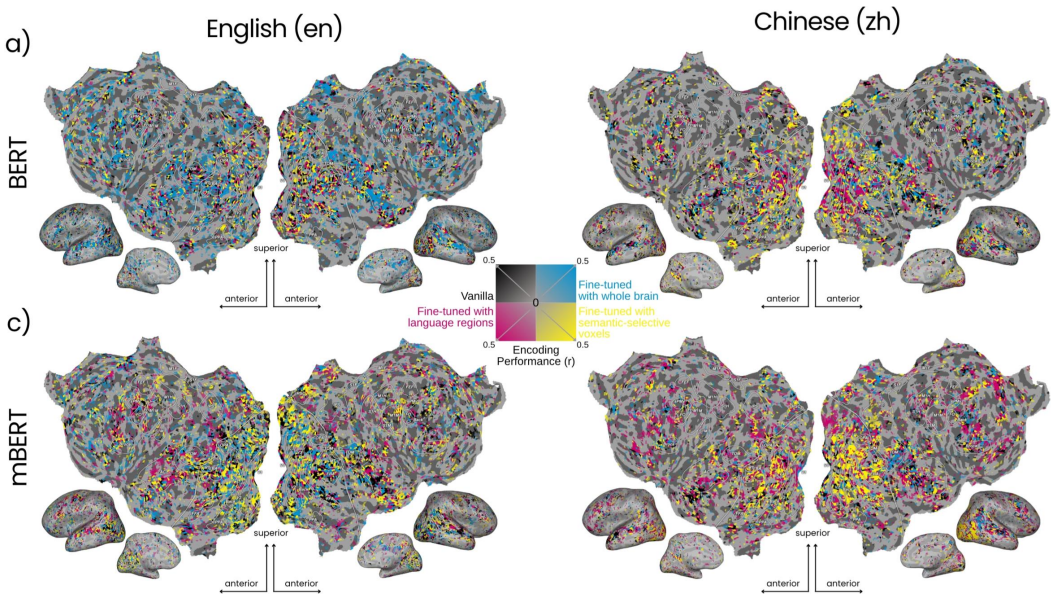

Figure 2: **Brain encoding performance before and after brain-informed fine-tuning.** We evaluated the effect of bilingual brain-informed fine-tuning using voxelwise encoding models (VEMs; see Methods). VEMs were trained using representations from both the vanilla (pretrained) model and their bilingual brain-informed fine-tuned variants (using whole-brain or using language, or semantically-selective brain regions). Flattened cortical surface for one participant is shown. Each voxel on the cortical surface is colored according to which model variant achieved the best encoding performance (r) for (a) BERT and (b) mBERT with English (left) and Chinese (right) brain data, respectively. Each point corresponds to a voxel that was consistently well predicted (r>0.1) across models. Voxel colors reflect the best-performing model: black for vanilla, blue for fine-tuned with whole-brain, magenta for fine-tuned with language and yellow for semantically-selective brain regions. Across both languages and model types, bilingual brain-informed fine-tuning consistently yields better encoding performance than its vanilla counterpart. Results for other participants are in Appendix D.

of the vanilla language model against the three fine-tuned variants, trained using whole-brain or language-selective, or subject-specific semantically-selective voxels. Fig. 2 shows flattened cortical surfaces for participant 1, visualizing which language model variant achieved the best encoding performance across well-predicted voxels (Pearson correlation r>0.1) for (a) monolingual and (b) multilingual models within the same language. Across both languages and model types, we found that fine-tuned language models outperform their vanilla counterparts, with the best-performing variant explaining more variance in a majority of voxels (76-84% of well-predicted voxels). This was consistent across all six participants (see Fig. 5 in Appendix E.2 for VEM results on other participants). Our results suggest that a language model's ability to predict brain activity benefits from brain-informed fine-tuning. Further, as shown in Appendix E.1, brain-informed fine-tuning outperforms control baselines (like TR-shuffled fMRI responses and mBERT representations as fine-tuning targets).

We do not observe any regional preference for any particular fine-tuning approach. No specific brain region consistently preferred one fine-tuned variant over the others across participants. Although fine-tuning improved encoding performance, the overall increase was modest (maximum $\Delta r \approx 0.15$), which aligns with prior studies. For example, previous work has shown that current representations from language models already serve as one of the best predictors of brain responses (Schrimpf et al., 2021), and even in speech models that lack brain-relevant semantics (Oota et al., 2023), fine-tuning yielded small improvements ($\Delta r \approx 0.1$) (Vattikonda et al., 2025).

To test whether these improvements were participant-specific, we checked for cross-participant generalization. We fine-tuned a BERT-en model on one participant's English brain data and evaluated its VEM performance on the other participants. As shown in Fig. 6 in Appendix E.3, we observe small but consistent improvements ($\Delta r \approx 0.03 - 0.05$) in high-level semantic areas. These findings

suggest that the improvements from brain-informed fine-tuning of language model representations are not specific to individual participants but reflect shared representations across bilingual individuals.

4.2 Performance on Downstream NLP tasks

We tested the effect of brain-informed fine-tuning on language models by evaluating model performance on standard NLP benchmarks. We assess this for three settings:

Fine-tuning and evaluation in the same language. We show in Table 1(a) performance on downstream NLP tasks when brain-informed fine-tuning and evaluation are computed in the same language. For English, models were fine-tuned using English brain data and evaluated on the GLUE benchmark. We observe that the monolingual model (BERT-ft-en) outperforms its vanilla counterpart (BERT-en) on 6/9 tasks, with an average improvement of +0.80 percentage points, and a maximum gain of +7.04 on the WNLI task. The multilingual model (mBERT-ft-en) improves on 7/9 tasks, with an average gain of +0.18 percentage points, and max gain of +2.82 also on WNLI. For Chinese, models were fine-tuned using Chinese brain data and evaluated on the CLUE benchmark. Here, BERT-ft-zh achieves slight improvements on 5/7 tasks relative to BERT-zh, with an average gain of +0.09 percentage point and max gain of +0.55 on the CSL task. The multilingual model (mBERT-ft-zh) also improves on 5/7 tasks, achieving an average improvement of +0.29 percentage point, with the largest improvement of +0.81 observed on the AFQMC task. These results indicate that brain-informed fine-tuning consistently improves downstream task performance for both monolingual and multilingual models when the fine-tuning and evaluation languages match.

Cross-language transfer between known languages. Table 1(b) reports performance when brain-informed fine-tuning is done in one language and evaluation is performed in the participant's other language. On the GLUE benchmark, we evaluated BERT-zh and mBERT models fine-tuned using Chinese brain data. Both models outperform their vanilla counterparts on all 9 tasks: BERT-ft-zh achieves an average improvement of +0.81 percentage points (max gain: +2.82 on WNLI), and mBERT-ft-zh shows an average improvement of +0.68 points (max gain: +1.41 on WNLI). On the CLUE benchmark, we evaluated BERT-en and mBERT models fine-tuned using English brain data. BERT-ft-en shows improvement on 5/7 tasks, with a slight decline in 2 tasks (-1.29 points on average). mBERT-ft-en improves across all 7 tasks, with an average gain of +0.27 points and a maximum improvement of +0.88 on the AFQMC task.

Overall, bilingual brain-informed fine-tuning enables cross-lingual generalization in language models, improving downstream task performance even when the fine-tuning and evaluation languages differ. This suggests that bilingual brain-informed fine-tuning of language models introduces changes that reflect the bilingual brain's shared semantic representations, enabling cross-language transfer.

Zero-shot transfer to unseen languages. Table 1(c) reports performance on downstream NLP tasks when brain-informed fine-tuning is performed with English brain data and evaluated on five unseen languages (German, French, Spanish, Japanese, and Korean), that were neither used for fine-tuning nor known to the participant. We evaluate the multilingual model (mBERT-ft-en) using tasks from the XGLUE and XTREME benchmarks. We observe that mBERT-ft-en improves performance on 2/3 tasks in German, French, Spanish, and Japanese, with average gains of +0.26, +1.61, +1.00, and +0.46 percentage points, respectively. In Korean, performance slightly declines with an average drop of -0.34 points. These results demonstrate that bilingual brain-informed fine-tuning improves language-agnostic representations in multilingual models, enabling generalization beyond the language of the brain data to entirely unseen languages (zero-shot).

Performance on downstream tasks with language- or semantically-selective voxels. We evaluated downstream task performance before and after brain-informed fine-tuning using only language- or semantically-selective voxels. Table 5 in Appendix F compares results for mBERT-en (fine-tuned with English brain data) and mBERT-zh (fine-tuned with Chinese brain data) when using whole-brain, language-selective, or semantic-selective regions for fine-tuning. Fine-tuning with either variant improves performance across all downstream tasks compared to vanilla models. Detailed analysis is provided in the Appendix. Results and analyses for other multilingual language models (XLM-R, XGLM, and LLaMA) are reported in the Supplementary.

Table 1: **Downstream task performance before and after bilingual brain-informed fine-tuning (with whole-brain).** We compared vanilla (pretrained) models, English BERT (BERT-en), Chinese BERT (BERT-zh), and mBERT, with their bilingual brain-informed fine-tuned counterparts (e.g., BERT-ft-en: BERT-en fine-tuned using English brain data). To assess whether brain-informed fine-tuning elicits multilingual capabilities, we evaluate downstream task performance in three settings: (a) Fine-tuning and evaluation in the same language: models are fine-tuned with brain data in one language (en or zh) and evaluated on NLP tasks in the same language (GLUE benchmark for en, CLUE benchmark for zh). This tests within-language improvements due to brain-informed fine-tuning. (b) Cross-language transfer between known languages: model is fine-tuned on brain data in one language and evaluated on tasks in the participants's second (not used in fine-tuning) language (e.g., fine-tuned with en brain data and evaluated on CLUE (zh benchmark) tasks). This tests whether bilingual brain-informed fine-tuning elicits the participants' shared semantic representations. (c) Zero-shot transfer to unseen languages: to assess broader multilingual transfer, mBERT-ft-en is evaluated on downstream tasks in additional languages not seen during fine-tuning (German, French, Spanish, Japanese, and Korean) using XGLUE and XTREME benchmarks. Bolded values indicate performance equal to or better than the corresponding vanilla model. We observe that bilingual brain-informed fine-tuning improves performance on several NLP tasks, with mBERT showing greater benefits than BERT across all three settings. Results for fine-tuning using language/semantically-selective regions are reported in Appendix F.1.

a) Fine-tuning and Evaluation in the Same Language

GLUE Task	BERT-en	BERT-ft-en	mBERT	mBERT-ft-en
CoLA	53.38	**55.75**	**42.68**	40.05
SST-2	92.08	**93.12**	89.68	**90.14**
MRPC (Acc.)	79.41	**79.66**	**84.80**	**84.80**
MRPC (F1)	**86.27**	86.19	88.56	**89.01**
STS-B (Pears.)	**88.06**	88.01	88.06	**88.42**
STS-B (Spear.)	**87.65**	87.63	87.76	**88.22**
QQP (Acc.)	**90.84**	90.77	90.22	**90.47**
QQP (F1)	**87.70**	87.60	86.70	**87.07**
MNLI-m	**84.38**	84.28	82.09	**82.66**
MNLI-mm	**84.64**	84.26	82.38	**82.97**
QNLI	91.45	**91.56**	91.14	**91.18**
RTE	67.15	**67.51**	**67.15**	65.70
WNLI	49.30	**56.34**	53.52	**56.34**

CLUE Task	BERT-zh	BERT-ft-zh	mBERT	mBERT-ft-zh
AFQMC	**75.25**	74.01	69.74	**70.55**
CMNLI	80.50	**80.89**	78.66	**79.02**
CSL (F1.)	80.18	**80.73**	**81.10**	80.71
IFLYTEK	60.25	**60.29**	56.52	**56.83**
TNEWS (Acc.)	56.24	**56.61**	54.77	**55.12**
TNEWS (F1.)	55.17	**55.52**	53.69	**53.74**
ChID	10.66	**10.66**	10.66	**10.66**
C^3	49.74	**49.97**	49.42	**50.21**

b) Cross-Language Transfer Between Known Languages

GLUE Task	BERT-zh	BERT-ft-zh	mBERT	mBERT-ft-zh
CoLA	51.13	**53.82**	40.96	**41.22**
SST-2	92.26	**92.85**	88.27	**89.25**
MRPC (Acc.)	76.96	**77.45**	82.84	**83.82**
MRPC (F1)	83.38	**84.20**	87.15	**87.88**
STS-B (Pears.)	87.21	**87.95**	87.09	**88.11**
STS-B (Spear.)	86.73	**87.65**	86.97	**87.83**
QQP (Acc.)	90.65	**90.76**	89.91	**90.16**
QQP (F1)	87.53	**87.72**	86.27	**86.89**
MNLI-m	84.00	**84.12**	81.52	**82.13**
MNLI-mm	83.91	**84.11**	81.94	**82.36**
QNLI	91.27	**91.38**	90.55	**90.92**
RTE	67.15	**67.87**	66.06	**66.43**
WNLI	53.52	**56.34**	54.93	**56.34**

CLUE Task	BERT-en	BERT-ft-en	mBERT	mBERT-ft-en
AFQMC	69.00	**69.00**	69.74	**70.62**
CMNLI	68.34	**68.63**	78.66	**78.83**
CSL (F1.)	71.20	**71.43**	81.10	**81.30**
IFLYTEK	**47.86**	46.79	56.52	**57.14**
TNEWS (Acc.)	50.92	**51.14**	54.77	**54.81**
TNEWS (F1.)	50.15	**50.60**	53.69	**53.84**
ChID	10.66	**10.66**	10.66	**10.66**
C^3	**42.64**	41.12	49.42	**49.48**

c) Zero-Shot Transfer to Unseen Languages

Task	German (de)		French (fr)		Spanish (es)		Japanese (ja)		Korean (ko)	
	mBERT	mBERT-ft-en	mBERT	mBERT-ft-en	mBERT	mBERT-ft-en	mBERT	mBERT-ft-en	mBERT	mBERT-ft-en
XGLUE										
PAWS-X	**84.00**	82.85	**88.10**	87.75	**87.05**	86.80	**78.70**	78.05	**78.75**	77.15
XNLI	70.60	**71.04**	72.69	**72.73**	74.10	**74.58**	86.00	**88.00**	86.00	**88.00**
NER	10.67	**13.00**	06.67	**12.33**	07.33	**12.67**	**13.00**	10.67	**12.00**	11.67
XTREME										
PAWS-X	**85.65**	83.90	**88.95**	87.30	**87.90**	87.25	**79.55**	77.95	**79.35**	78.25
XNLI	70.06	**71.08**	**72.97**	72.25	74.50	**74.60**	86.00	**87.00**	86.00	**87.00**
NER	12.33	**13.00**	09.33	**14.00**	10.33	**11.33**	08.67	**13.00**	**12.33**	10.33

4.3 Monolingual Brain data as Fine-Tuning Targets

Table 2 compares downstream task performance after brain-informed fine-tuning with English brain data from either bilingual or monolingual participant. On GLUE benchmark (within-language evaluation), bilingual tuning outperforms monolingual tuning on 5/7 tasks, especially in all inference-related tasks (MNLI, QNLI, WNLI, and RTE). On CLUE benchmark (cross-language evaluation),

Table 2: **Downstream task performance before and after bilingual or monolingual brain-informed fine-tuning.** We perform brain-informed fine-tuning of mBERT with English whole-brain brain data (mBERT-ft-en) from either a bilingual (participant 1) or a monolingual participant. We evaluate downstream task performance in two settings: (a) Fine-tuning and evaluation in the same language: the model is evaluated in English with GLUE tasks. (b) Cross-language transfer between known languages: the model is evaluated in Chinese with CLUE tasks.

a) Fine-tuning and Evaluation in the Same Language

		mBERT-ft-en	
Task	**mBERT**	**Bilingual**	**Monolingual**
CoLA	**42.68**	40.05	35.74
SST-2	89.68	90.14	**90.48**
MRPC (Acc.)	84.80	**84.80**	78.85
MRPC (F1)	88.56	**89.01**	85.78
STS-B (Pearson)	88.06	**88.42**	**88.42**
STS-B (Spearman)	87.76	88.22	**88.34**
QQP (Acc.)	90.22	90.47	**90.54**
QQP (F1)	86.70	87.07	**87.19**
MNLI-m	82.09	**82.66**	82.41
MNLI-mm	82.38	**82.97**	82.78
QNLI	91.14	**91.18**	91.12
RTE	67.15	**65.70**	65.34
WNLI	53.52	**56.34**	**56.34**

b) Cross-Language Transfer Between Known Languages

		mBERT-ft-en	
CLUE Task	**mBERT**	**Bilingual**	**Monolingual**
AFQMC	69.74	70.62	**70.64**
CMNLI	78.66	**78.83**	78.48
CSL (F1)	81.10	**81.30**	81.18
IFLYTEK	46.79	56.52	**57.06**
TNEWS (Acc.)	54.77	**54.81**	54.80
TNEWS (F1)	53.69	**53.84**	53.50
ChID	10.66	**10.66**	**10.66**
C^3	49.42	**49.48**	49.46

bilingual brain-informed fine-tuning leads to higher performance on 5/7 tasks, while monolingual tuning leads on 2. These results suggest that while monolingual brain-informed fine-tuning also benefits downstream tasks within-language, fine-tuning with bilingual brain data enhances cross-linguistic generalization.

5 Discussion and Conclusion

In this work, we perform brain-informed fine-tuning of monolingual and multilingual language models using brain data from bilingual participants reading the same naturalistic stories in English and Chinese. We show that bilingual brain-informed fine-tuned language models outperform their vanilla (pretrained) counterparts in both brain encoding performance and most downstream tasks. Our analysis of brain-informed fine-tuning reveals several key conclusions.

First, representations from brain-informed fine-tuned language models are shared across bilingual individuals. Both monolingual and multilingual language models show improved encoding performance after brain-informed fine-tuning, in both English and Chinese (Fig.2). These models outperform baseline controls (Appendix Fig.4) and generalize across participants (Appendix Fig.6). This highlights the potential of bilingual brain data as a rich supervisory signal for grounding language model representations with biologically meaningful representations. Second, brain-informed fine-tuning enables cross-language transfer in language models. Downstream performance improves across within-language, cross-language, and unseen language settings (Table 1). This holds for fine-tuning with whole-brain or language-selective or semantically-selective regions (Appendix Table 5). This effect is likely driven by shared semantic representations in bilingual individuals (Chen et al., 2024b). This highlights the potential of leveraging bilingual brain representations for developing language-agnostic language models. Lastly, the observed improvements are driven specifically by brain-informed fine-tuning with bilingual brain data, not brain data in general. Fine-tuning with monolingual brain data does not yield similar cross-linguistic transfer (Table 2).

Our results contribute to a growing body of work exploring the representational alignment between brain and artificial language models. We demonstrate that bilingual brain-informed fine-tuning improves multilingual understanding in language models. To the best of our knowledge, this is the first work to do so. This study offers a bridge between neuroscience research on cross-linguistic semantic representation and NLP efforts to interpret and improve multilingual language models.

6 Limitations and Future Work

A key limitation of this study is the relatively small number of data samples per subject, due to limited bilingual brain recordings with naturalistic stimuli. This constrains the observed performance gains. Additionally, our analysis is restricted to only two languages (English and Chinese). Expanding to

larger and more diverse multilingual brain datasets would allow a broader evaluation of cross-linguistic transfer. Future work could also examine what linguistic properties are captures by brain-informed fine-tuning (such as syntax, morphology, or discourse-level structure). This can potentially be tested by targeting functionally selective brain regions and analyzing fine-tuned model representations across layers.

Acknowledgment

We thank Lily Gong for preprocessing the fMRI data, 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

Mostafa Abdou, Ana Valeria González, Mariya Toneva, Daniel Hershcovich, and Anders Søgaard. Does injecting linguistic structure into language models lead to better alignment with brain recordings? *arXiv preprint arXiv:2101.12608*, 2021.

Richard Antonello, Javier S Turek, Vy Vo, and Alexander Huth. Low-dimensional structure in the space of language representations is reflected in brain responses. *Advances in Neural Information Processing Systems*, 34:8332–8344, 2021.

Khai Loong Aw and Mariya Toneva. Training language models to summarize narratives improves brain alignment. In *The Eleventh International Conference on Learning Representations*, 2023.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*, 2016.

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

Catherine Chen, Tom Dupré la Tour, Jack L Gallant, Daniel Klein, and Fatma Deniz. The cortical representation of language timescales is shared between reading and listening. *Communications Biology*, 7(1):284, 2024a.

Catherine Chen, Xue L Gong, Christine Tseng, Daniel L Klein, Jack L Gallant, and Fatma Deniz. Bilingual language processing relies on shared semantic representations that are modulated by each language. *bioRxiv*, pp. 2024–06, 2024b.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 8440–8451, 2020.

Andrea Gregor de Varda, Saima Malik-Moraleda, Greta Tuckute, and Evelina Fedorenko. Multilingual computational models reveal shared brain responses to 21 languages. *bioRxiv*, pp. 2025–02, 2025.

Fatma Deniz, Anwar O Nunez-Elizalde, Alexander G Huth, and Jack L Gallant. The representation of semantic information across human cerebral cortex during listening versus reading is invariant to stimulus modality. *Journal of Neuroscience*, 2019.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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, 2019.

Evelina Fedorenko, Po-Jang Hsieh, Alfonso Nieto-Castañón, Susan Whitfield-Gabrieli, and Nancy Kanwisher. New method for fmri investigations of language: defining rois functionally in individual subjects. *Journal of neurophysiology*, 104(2):1177–1194, 2010.

Wendy S Francis. Bilingual semantic and conceptual representation. *Handbook of bilingualism: Psycholinguistic approaches*, pp. 251–267, 2005.

Ariel Goldstein, Zaid Zada, Eliav Buchnik, Mariano Schain, Amy Price, Bobbi Aubrey, Samuel A Nastase, Amir Feder, Dotan Emanuel, Alon Cohen, et al. Shared computational principles for language processing in humans and deep language models. *Nature neuroscience*, 25(3):369–380, 2022.

François Grosjean. *On bilinguals and bilingualism*. Cambridge University Press, 2024.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *International conference on machine learning*, pp. 4411–4421. PMLR, 2020.

Alexander G Huth, Wendy A De Heer, Thomas L Griffiths, Frédéric E Theunissen, and Jack L Gallant. Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, 532 (7600):453–458, 2016.

Shailee Jain and Alexander Huth. Incorporating context into language encoding models for fmri. *Advances in Neural Information Processing Systems*, 31, 2018.

Antonia Karamolegkou, Mostafa Abdou, and Anders Søgaard. Mapping brains with language models: A survey. *arXiv preprint arXiv:2306.05126*, 2023.

Mathis Lamarre, Catherine Chen, and Fatma Deniz. Attention weights accurately predict language representations in the brain. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 4513–4529, 2022.

Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, et al. Xglue: A new benchmark datasetfor cross-lingual pre-training, understanding and generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6008–6018. Association for Computational Linguistics, 2020.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona Diab, Veselin Stoyanov, and Xian Li. Few-shot learning with multilingual generative language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 9019–9052, 2022.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Omer Moussa, Dietrich Klakow, and Mariya Toneva. Improving semantic understanding in speech language models via brain-tuning. In *The Thirteenth International Conference on Learning Representations*, 2025. URL `https://openreview.net/forum?id=KL8Sm4xRn7`.

Thomas Naselaris, Kendrick N Kay, Shinji Nishimoto, and Jack L Gallant. Encoding and decoding in fmri. *Neuroimage*, 56(2):400–410, 2011.

Anwar O Nunez-Elizalde, Alexander G Huth, and Jack L Gallant. Voxelwise encoding models with non-spherical multivariate normal priors. *Neuroimage*, 197:482–492, 2019.

Subba Reddy Oota, Jashn Arora, Veeral Agarwal, Mounika Marreddy, Manish Gupta, and Bapi Surampudi. Neural language taskonomy: Which nlp tasks are the most predictive of fmri brain activity? In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3220–3237, 2022.

Subba Reddy Oota, Agarwal Veeral, Marreddy Mounika, Gupta Manish, and Raju Surampudi Bapi. Speech taskonomy: Which speech tasks are the most predictive of fmri brain activity? In *24th INTERSPEECH Conference*, 2023.

Subba Reddy Oota, Emin Çelik, Fatma Deniz, and Mariya Toneva. Speech language models lack important brain-relevant semantics. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8503–8528. Association for Computational Linguistics, 2024a.

Subba Reddy Oota, Manish Gupta, and Mariya Toneva. Joint processing of linguistic properties in brains and language models. *Advances in Neural Information Processing Systems*, 36, 2024b.

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

Lisa Schut, Yarin Gal, and Sebastian Farquhar. Do multilingual llms think in english? *arXiv preprint arXiv:2502.15603*, 2025.

Dan Schwartz, Mariya Toneva, and Leila Wehbe. Inducing brain-relevant bias in natural language processing models. *Advances in Neural Information Processing Systems*, 32, 2019.

Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. *Advances in Neural Information Processing Systems*, 29, 2016.

Mariya Toneva and Leila Wehbe. Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain). *Advances in Neural Information Processing Systems*, 32, 2019.

Mariya Toneva, Jennifer Williams, Anand Bollu, Christoph Dann, and Leila Wehbe. Same cause; different effects in the brain. *Causal Learning and Reasoning*, 2022.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Nishitha Vattikonda, Aditya R Vaidya, Richard J Antonello, and Alexander G Huth. Brainwavlm: Fine-tuning speech representations with brain responses to language. *arXiv preprint arXiv:2502.08866*, 2025.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *International Conference on Learning Representations*, 2018.

Leila Wehbe, Brian Murphy, Partha Talukdar, Alona Fyshe, Aaditya Ramdas, and Tom Mitchell. Simultaneously uncovering the patterns of brain regions involved in different story reading subprocesses. *PloS One*, 11, 2014a.

Leila Wehbe, Ashish Vaswani, Kevin Knight, and Tom Mitchell. 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, 2014b.

Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. Do llamas work in english? on the latent language of multilingual transformers. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15366–15394, 2024.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 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, 2020.

Michael C-K Wu, Stephen V David, and Jack L Gallant. Complete functional characterization of sensory neurons by system identification. *Annual Review of Neuroscience*, 29(1):477–505, 2006.

Liang Xu, Hai Hu, Xuanwei Zhang, Lu Li, Chenjie Cao, Yudong Li, Yechen Xu, Kai Sun, Dian Yu, Cong Yu, et al. Clue: A chinese language understanding evaluation benchmark. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 4762–4772, 2020.

Overview of Appendices

- Section A: Detailed Related Work
- Section B: Naturalistic Brain Imaging Dataset and Preprocessing
- Section C: Brain-Informed Fine-tuning with Subsets of Cortex
- Section D: Details of Downstream NLP Tasks
- Section E: Additional results for Voxelwise encoding performance
- Section F Models fine-tuned with different training objectives
- Section G: Implementation details for reproducibility

A Detailed Related Work

Fine-tuning of language models with naturalistic brain data. Our work is most closely related to that of Schwartz et al. (2019); Moussa et al. (2025); Vattikonda et al. (2025), who proposed the brain-tuning approach to fine-tune pretrained Transformer-based language models using brain data. This approach integrates brain-relevant information into the language models and examines whether and how brain-informed language models effect brain encoding performance and downstream task performance. Specifically, Schwartz et al. (2019) introduced brain-tuning using brain data recorded during naturalistic reading tasks in monolingual English speakers, observing improved brain encoding performance after fine-tuning along with enhanced performance on downstream NLP tasks. More recently, Moussa et al. (2025); Vattikonda et al. (2025) applied a similar approach to explore whether brain-tuning of speech-based language models could enhance brain-relevant semantic representations, thus improving encoding performance and speech-task performance. Our study complements these previous studies by investigating bilingual brain-informed fine-tuning and exploring how monolingual and multilingual language models change when trained with bilingual brain data.

Multilingual language models and brain alignment. Our work also relates to a growing body of literature investigating alignment between human brain activity and language models. Several studies have successfully used text-based language models to predict brain activity evoked by both written and spoken stimuli, achieving impressive levels of alignment between language models and brain activity (Wehbe et al., 2014a; Jain & Huth, 2018; Toneva & Wehbe, 2019; Deniz et al., 2019; Abdou et al., 2021; Toneva et al., 2022; Antonello et al., 2021; Oota et al., 2022; Aw & Toneva, 2023; Oota et al., 2024b; Lamarre et al., 2022; Chen et al., 2024a). More recently, research has focused on multilingual Transformer-based language models using brain data from reading and listening tasks across multiple languages to evaluate their brain encoding performance (de Varda et al., 2025). However, previous multilingual studies have generally remained monolingual in their experimental design, with participants exposed exclusively to stimuli in one language. A notable exception is the recent work by Chen et al. (2024a), which examined bilingual language processing in participants who read identical stories in both English and Chinese, and find that bilingual individuals have shared semantic representations across languages. Our approach complements this research by further exploring bilingual brain alignment through brain-informed fine-tuning of language models, particularly analyzing how different brain regions respond to bilingual stimuli and fine-tuning strategies, evaluating their impact on downstream NLP tasks in both English and Chinese.

B Naturalistic Brain Imaging Dataset and Preprocessing

B.1 MRI data collection

MRI data were collected on a 3T Siemens TIM Trio scanner using a 32-channel volume coil. Functional scans used a gradient-echo EPI sequence (TR = 2.0045 s, TE = 35 ms, flip angle = 74°, voxel size = 2.24 × 2.24 × 4.1 mm, 30 interleaved axial slices). Motion correction and alignment were performed using the FMRIB Linear Image Registration Tool (FLIRT) from FSL.

B.2 Participants

Functional data was collected from six healthy bilingual (in English and Chinese) and one healthy monolingual (English native) adult(s) with normal hearing and vision. All subjects were right-handed

or ambidextrous. All potential risks and/or discomforts from MRI were disclosed to participants: 1) loud beeping and hammering sounds while the scanner is collecting measurements; 2) claustrophobia while inside the scanner; 3) nerve stimulation; 4) risk of injury from metal objects affected by the scanner (highly unlikely); 5) compromised confidentiality (highly unlikely). All subjects signed a consent form where these risks were disclosed. All ethical regulations relevant to human research participants were followed, and all procedures were approved by the Committee for the Protection of Human Subjects. For anonymity, detailed information about the ethics committee has been omitted and will be included in the camera-ready version. Voxelwise encoding models and brain-informed fine-tuning are done separately for each subject, with results reported per subject. No sample size calculations were performed, as each subject serves as a full replication of the results.

C Brain-Informed Fine-tuning with Subsets of Cortex

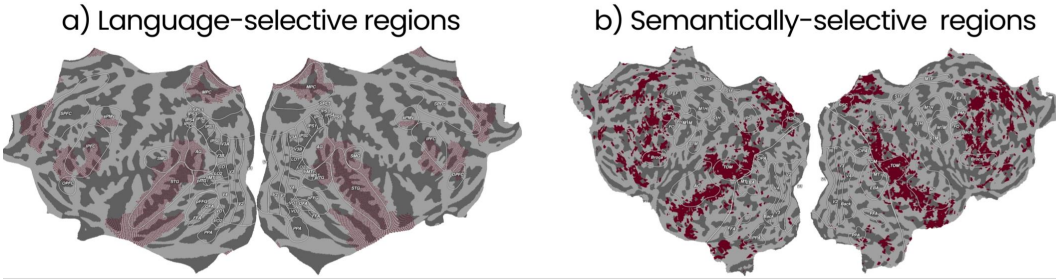

Figure 3: **Cortical regions used for brain-informed fine-tuning.** Flattened cortical surfaces for: (a) Language-selective regions displayed on the 'fsaverage' surface, used as the mask for all participants. and (b) semantically-selective regions for English for Participant 1. This mask is computed separately for each participant and language. Only fMRI responses from the dark red regions were used for fine-tuning in their respective variants.

C.1 Fine-tuning with Language-Selective Regions

Voxels were extracted from regions of interest (ROIs) known to support language comprehension (Fedorenko et al., 2010). To apply these ROIs to individual subjects, we first mapped each participant's cortical surface to the fsaverage template using surface-based registration. The group-level language mask was then projected onto each subject's brain to extract subject-specific language ROIs. The ROI mask used is shown in Appendix Fig. 3a.

C.2 Fine-tuning with Semantically-Selective Regions

To identify voxels selective for semantics, we applied a two-stage regression procedure to isolate semantic signals while controlling for low-level sensory confounds. Semantic features were extracted using fastText (Bojanowski et al., 2016) embeddings of the narrative text. This procedure was performed separately for each subject and language. First, we trained VEMs for response to seven low-level sensory feature spaces: word count, letter/character count, single phonemes, diphones, triphones, visual motion energy, and pixel-based orthographic similarity. These models were trained on the training data and used to predict responses for both train and test sets. Predicted low-level responses were subtracted from the actual BOLD signals to obtain residuals. These residual responses were then z-scored per voxel to have unit variance. Next, we trained VEMs using the semantic features on the residual signals. Voxels that showed significant (one-sided permutation test, $p < 0.05$, FDR-corrected) prediction accuracy based on semantic features were labeled as semantic-selective. The semantic-selective voxel mask for participant 1 is shown in Appendix Fig. 3b.

C.3 Baseline Models for Fine-tuning

TR-shuffled fMRI as fine-tuning target. We tested how fine-tuning language models with temporally misaligned data affects model representation. This TR-shuffled approach allows us to verify that the improvements from brain-informed fine-tuning are driven by meaningful stimulus-response

alignment, rather than random brain data. Here, we randomly shuffled the brain responses, permuted in blocks of 10 contiguous TRs. This breaks the correspondence between stimulus and brain response.

Multilingual model features as fine-tuning target. We tested how fine-tuning monolingual language models with multilingual model representations affects model representations. The goal of this test is to compare the results between pure-language model fine-tuning with brain-informed language model fine-tuning. Here, we replaced the fMRI data associated with input stimuli with representations obtained from a multilingual language model. We then fine-tuned monolingual language models (BERT-en and BERT-zh) using these multilingual model representations (from mBERT) as targets.

Monolingual brain data as fine-tuning target. We tested how tuning language models with brain data from individuals who know only one language differs from tuning language models with bilingual brain data. This test is used to clarify the impact of bilingualism on the resulting model representations.

D Details of Downstream NLP Tasks

Detailed description of the downstream NLP tasks in GLUE and CLUE benchmarks is provided in Table 3. Detailed description of the downstream NLP tasks in XGLUE and XTREME benchmarks is provided in Table 4.

Table 3: Descriptions of tasks in the GLUE and CLUE benchmarks.

Task	Description	Task Type
	GLUE Benchmark	
CoLA	Determine linguistic acceptability of a sentence	Single-sentence classification
SST-2	Sentiment analysis on movie reviews	Single-sentence classification
MRPC	Determine if two sentences are semantically equivalent	Sentence-pair classification
STS-B	Measure semantic similarity (0–5) between two sentences	Sentence-pair regression
QQP	Identify if two questions are duplicates	Sentence-pair classification
MNLI	Determine if a premise entails, contradicts, or is neutral to a hypothesis	Sentence-pair classification (multi-class)
QNLI	Determine if a sentence contains the answer to a question	Sentence-pair classification
RTE	Determine textual entailment between two sentences	Sentence-pair classification (binary)
WNLI	Coreference resolution: determine if a hypothesis is true given a sentence with a replaced pronoun	Sentence-pair classification
	CLUE Benchmark	
AFQMC	Determine if two Chinese sentences are semantically equivalent	Sentence-pair classification
CMNLI	Multilingual version of MNLI for Chinese	Sentence-pair classification
CSL	Determine whether a keyword list matches a given abstract	Sentence-pair classification
IFLYTEK	Classify app descriptions into predefined categories	Multi-class classification
TNEWS	Classify news headlines into categories	Multi-class classification
ChID	Fill-in-the-blank multiple-choice cloze test from Chinese literature	Cloze-style multiple-choice
C^3	Reading comprehension with multiple-choice answers	Multi-turn QA / Multiple-choice

Table 4: Summary of XGLUE and XTREME tasks used in this study.

Benchmark	Task Name	Task Type	Languages Evaluated	Evaluation Metric
XGLUE	PAWS-X	Paraphrase Identification	de, fr, es, ja, ko	Accuracy
	XNLI	Natural Language Inference	de, fr, es, ja, ko	Accuracy
	NER	Named Entity Recognition	de, fr, es, ja, ko	F1 Score
XTREME	PAWS-X	Paraphrase Identification	de, fr, es, ja, ko	Accuracy
	XNLI	Natural Language Inference	de, fr, es, ja, ko	Accuracy
	NER	Named Entity Recognition	de, fr, es, ja, ko	F1 Score

E Additional results for Voxelwise Encoding Performance

Please note that we used embeddings extracted from the language model layer for fitting the VEMs. No low-level sensory features were included. As a result, well-predicted voxels in reported figures do not necessarily indicate language-selective brain regions. This distinction is important to avoid overinterpreting encoding performance in non-linguistic areas such as the early visual cortex.

E.1 Baseline Models

Appendix Fig. 4 shows results from baseline experiments where brain-informed fine-tuning was performed using (a) TR-shuffled fMRI responses and (b) mBERT representations as fine-tuning targets. In both cases, we observe that VEM predictions are limited to some sensory regions and scattered across the cortex, with no systematic predictions in higher-level semantic areas. This

confirms that the observed improvements in our main experiments are not driven by spurious correlations but rather due to biologically plausible representations in bilingual brains.

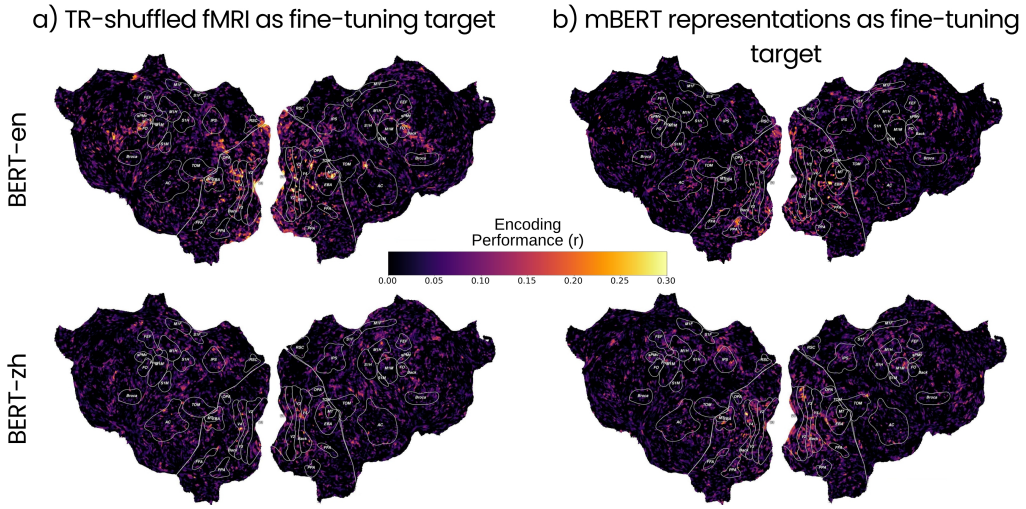

Figure 4: **Encoding performance of baseline fine-tuned models.** Voxelwise encoding performance is visualised for two fine-tuning baselines: (a) using TR-shuffled fMRI data, and (b) using mBERT representations as fine-tuning target. Flattened cortical surfaces with encoding performance are shown for Participant 1, with BERT-en (on top) and BERT-zh (on the bottom) fine-tuned with the baseline targets. In both baselines, the well predicted voxels do not correspond to semantically regions and appear scattered or in low-level sensory regions.

E.2 Results on Participants 2-6

Fig. 5 shows brain encoding performance before and after brain-informed fine-tuning for all participants. This figure shows voxelwise encoding model performance for the best-performing fine-tuned variants across the cortex for all participants, using (a) BERT-en, (b) BERT-zh, and (c–d) mBERT models. This replicates the format of Fig. 2, and reports it for all participants. For details on voxel color coding, refer to the caption of Fig. 2. Across participants, model types, and languages, brain-informed fine-tuning consistently improves encoding performance relative to it's vanilla counterpart.

E.3 Cross-participant Transfer

To examine whether brain-informed fine-tuning generalizes encoding performance across participants, we evaluated brain-informed fine-tuned models on one participant's brain data and test it on the brain recordings of other participants. Fig. 6 presents the encoding performance comparisons between vanilla and brain-tuned models in this cross-participant setting. We observe that there is also no difference in overall encoding performance between the vanilla and brain-tuned models when tested on unseen participants. This pattern holds consistently across all participants, for both English and Chinese brain data, and for both monolingual and multilingual BERT models. These findings suggest that the representational changes introduced by brain-specific fine-tuning are not participant-specific but reflect some shared representations across bilingual individuals.

F Additional results on Downstream NLP Tasks Performance

F.1 Vanilla vs. Fine-tuned with Whole-Brain/Language-Selective regions/Semantic-Selective voxels

Table 5 shows downstream task performance before and after bilingual brain-informed fine-tuning (with whole-brain, language-selective, and semantic-selective voxels). We compared vanilla (pre-trained) models and mBERT, with their bilingual brain-informed fine-tuned counterparts (e.g., mBERT-ft-en and mBERT-ft-zh). To assess whether brain-informed fine-tuning elicit multilingual capabilities, we evaluate downstream task performance in two settings: (a) Fine-tuning and

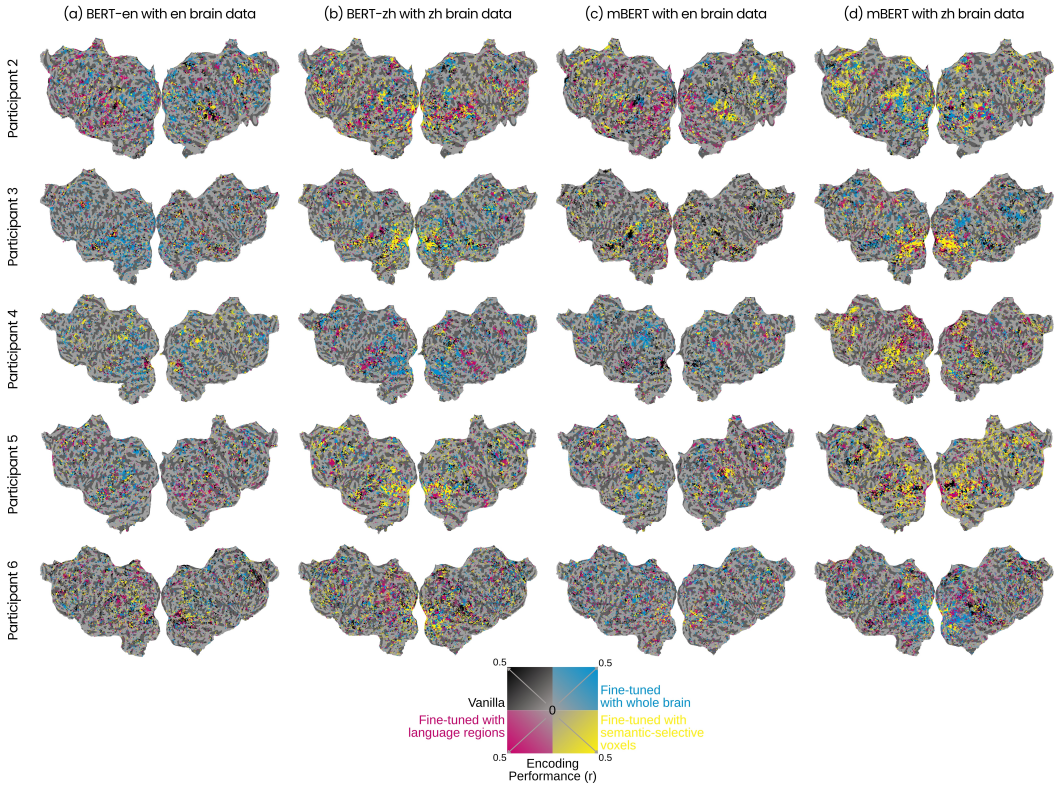

Figure 5: **Brain encoding performance before and after brain-informed fine-tuning for all participants.** This figure shows voxelwise encoding model performance for the best-performing fine-tuned variants across the cortex for all participants, using (a) BERT-en, (b) BERT-zh, and (c–d) mBERT models. This replicates the format of Fig. 2, and reports it for all participants. For details on voxel color coding, refer to the caption of Fig. 2. Across participants, model types, and languages, brain-informed fine-tuning consistently improves encoding performance relative to it's vanilla counterpart.

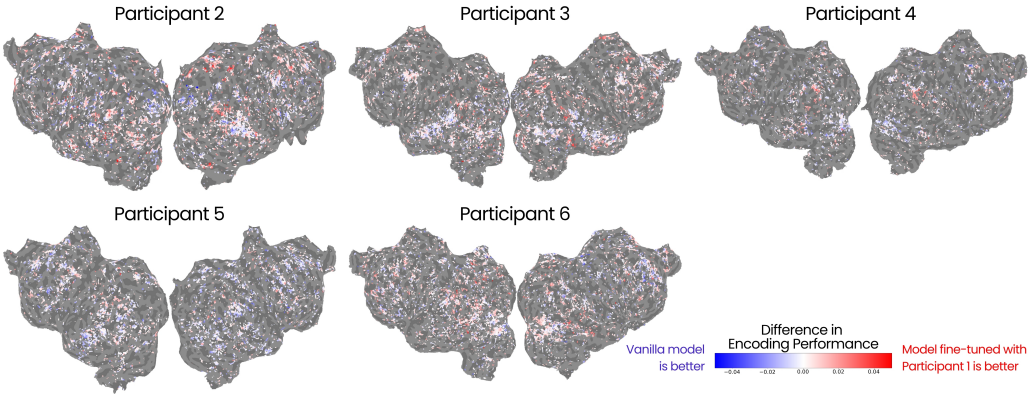

Figure 6: **Cross-participant generalization of brain-informed fine-tuned language models.** Flattened cortical surfaces show changes in encoding performance for Participants 2–6 when comparing BERT-en (vanilla) with BERT-ft-en fine-tuned using English brain data from Participant 1. Each point represents a voxel that was consistently well-predicted (r>0.1) across both models. Across participants, encoding performance in higher-level semantic regions shows either no degradation or slight improvements with the fine-tuned model.

Evaluation in the Same Language: models are fine-tuned with brain data in one language (English (en) or Chinese (zh)) and evaluated on NLP tasks in the same language (GLUE benchmark for

en, CLUE benchmark for zh). This tests within-language improvements due to brain-informed fine-tuning. (b) Cross-Language Transfer Between Known Languages: model is fine-tuned on brain data in one language and evaluated on tasks in the participants's second (not used in fine-tuning) language (e.g., fine-tuned with en brain data and evaluated on CLUE (zh benchmark) task. This tests whether bilingual brain-informed fine-tuning elicits the participants' shared semantic representations. Bolded values indicate performance equal to or better than the corresponding vanilla model. We observe that bilingual brain-informed fine-tuning improves performance on several NLP tasks, with either language- or semantic-selective region variants showing more improvement than their vanilla counterpart.

G Implementation details for reproducibility

All brain-informed fine-tuning experiments were conducted on a machine equipped with an NVIDIA TITAN RTX GPU (24 GB RAM), and NVIDIA RTX A6000 GPU (40GB RAM). The downstream NLP tasks were evaluated on machines with the same GPU configuration. Voxelwise encoding models were trained on the TITAN RTX GPU using banded ridge regression, with the following hyperparameters: MSE loss, L2 regularization (λ) values logarithmically spaced from 10^{-10} to 10^{10}, target batch size of 1000, and 20 λ values per batch.

For downstream NLP task fine-tuning, we used a batch size of 64, learning rate of 2e-5, weight decay of 0.01, per-device evaluation batch size of 128, and trained for 3 epochs.

Table 5: **Downstream task performance before and after bilingual brain-informed fine-tuning (with whole-brain, or language-selective, or semantically-selective regions).** We compared vanilla (pretrained) models and mBERT, with their bilingual brain-informed fine-tuned counterparts (e.g., mBERT-ft-en and mBERT-ft-zh). To assess whether brain-informed fine-tuning elicit multilingual capabilities, we evaluate downstream task performance in two settings: (a) Fine-tuning and evaluation in the same language: models are fine-tuned with brain data in one language (English (en) or Chinese (zh)) and evaluated on NLP tasks in the same language (GLUE benchmark for en, CLUE benchmark for zh). This tests within-language improvements due to brain-informed fine-tuning. (b) Cross-language transfer between known languages: model is fine-tuned on brain data in one language and evaluated on tasks in the participants' second (not used in fine-tuning) language (e.g., fine-tuned with en brain data and evaluated on CLUE (zh benchmark) task. This tests whether bilingual brain-informed fine-tuning elicits the participants' shared semantic representations. Bolded values indicate performance equal to or better than the corresponding vanilla model. We observe that bilingual brain-informed fine-tuning improves performance on several NLP tasks, with either language-selective or semantic-selective voxels showing greater benefits vanilla models.

a) Fine-tuning and Evaluation in the Same Language

Task	Vanilla	Whole-Brain	Language	Semantic
QNLI	91.14	91.18	91.10	**91.31**
CoLA	42.68	40.05	41.80	**42.81**
MNLI-mm	82.38	**82.97**	82.92	82.68
MNLI-m	82.09	**82.66**	82.27	82.07
QQP (Acc)	90.22	90.47	90.43	**90.48**
QQP (F1)	86.70	87.07	87.02	**87.11**
STS-B (P)	88.06	88.42	**88.46**	88.42
STS-B (SP)	87.76	88.22	**88.38**	88.34
MRPC (F1)	88.56	89.01	**90.49**	89.73
MRPC (Acc)	84.80	84.80	**86.76**	86.03
SST-2	89.68	90.14	90.25	**90.60**
WNLI	53.52	**56.34**	**56.34**	**56.34**
RTE	67.15	65.70	**68.95**	67.87

CLUE Task	mBERT-zh	Whole-Brain	Language	Semantic
AFQMC	69.74	70.55	**72.03**	71.11
CMNLI	78.66	**79.02**	78.75	79.01
CSL (F1.)	81.10	80.71	**82.27**	81.24
IFLYTEK	56.52	56.83	56.79	**56.91**
TNEWS (Acc.)	54.77	55.12	**55.27**	54.91
TNEWS (F1.)	53.69	53.74	53.82	**53.98**
ChID	10.66	**10.66**	10.66	**10.66**
C^3	49.42	**50.21**	49.76	49.95

b) Cross-Language Transfer Between Known Languages

Task	Vanilla	Whole-Brain	Language	Semantic
QNLI	90.55	90.92	91.27	**91.38**
CoLA	40.96	41.22	40.82	**41.60**
MNLI-mm	81.94	82.36	82.74	**83.12**
MNLI-m	81.52	82.13	82.54	**82.94**
QQP (Acc)	89.91	90.16	**90.46**	90.42
QQP (F1)	86.27	86.89	**87.06**	87.00
STS-B (P)	87.09	88.11	87.00	**88.54**
STS-B (SP)	86.97	87.83	87.27	**88.43**
MRPC (F1)	87.15	87.88	89.09	**90.62**
MRPC (Acc)	82.84	83.82	85.05	**87.01**
SST-2	88.27	89.25	**91.17**	90.37
WNLI	54.93	**56.34**	**56.34**	**56.34**
RTE	66.06	66.43	**68.59**	66.06

CLUE Task	mBERT-en	Whole-Brain	Language	Semantic
AFQMC	69.74	70.62	**71.29**	70.76
CMNLI	78.66	78.83	78.92	**79.09**
CSL (F1.)	81.10	81.30	81.53	**81.57**
IFLYTEK	46.79	56.52	**56.60**	56.33
TNEWS (Acc.)	54.77	54.81	**54.95**	53.47
TNEWS (F1.)	53.69	53.84	**54.01**	53.75
ChID	10.66	**10.66**	10.66	**10.66**
C^3	49.42	49.48	**49.84**	49.76