What Brain Data Adds to Language Model Training

Gabriele Merlin
MPI-SWS
gmerlin@mpi-sws.org

Omer Moussa
MPI-SWS
omoussa@mpi-sws.org

Mariya Toneva
MPI-SWS
mtoneva@mpi-sws.org

Abstract

Brain-tuning language models (LMs)—fine-tuning LMs to predict brain recordings elicited by linguistic stimuli—has been proposed as a promising way to align LMs closer to the human brain, with recent work reporting gains on a small number of downstream tasks. However, it remains unclear what benefits brain data provide beyond those obtainable from further training on the same underlying linguistic input, and whether such benefits generalize across tasks. Here, we present a comprehensive evaluation of Jointly-Tuned LMs, trained on both brain recordings and text-based stimuli, Brain-Tuned LMs and LMs tuned only on text-based stimuli (i.e., Stimulus-Tuned LMs). We compare models across a diverse suite of downstream linguistic tasks. We find that Jointly-Tuned LMs outperform other fine-tuned and pretrained models, and that Brain-Tuned LMs outperform Stimulus-Tuned LMs, demonstrating the richness of brain data as an additional training signal for LMs.

1 Introduction

Language models (LMs) have achieved remarkable success by scaling training on vast quantities of text, yet they remain imperfect models of human language understanding. Motivated by the idea that neural responses reflect constraints and inductive biases of the human language system, a growing body of work has begun to align LMs closer to the human brain via brain-tuning: fine-tuning LMs to predict human brain recordings elicited by linguistic stimuli (Moussa et al., 2025; Vattikonda et al., 2025; Freteault et al., 2025; Negi et al., 2025; Schwartz et al., 2019). Early results suggest that incorporating fMRI data during training can improve alignment of speech-based models (Moussa et al., 2025; Vattikonda et al., 2025; Moussa and Toneva, 2025) and text-based models (Schwartz et al., 2019; Negi et al., 2025) with neural measurements and yield gains on select downstream speech (Moussa et al., 2025; Freteault et al., 2025; Moussa and Toneva, 2025) and text (Negi et al., 2025) tasks, raising the possibility that brain signals provide a training signal that goes beyond speech or text alone.

However, a fundamental question remains unresolved: what does brain data add to LMs? In nearly all prior work, Brain-Tuned models are compared to baselines that differ not only in their supervision signal, but also in their exposure to linguistic input (Vattikonda et al., 2025; Freteault et al., 2025; Negi et al., 2025; Schwartz et al., 2019). This leaves open a critical alternative explanation: namely, that observed gains arise from additional training on the stimulus text itself, rather than from information uniquely contributed by neural responses. Moreover, reported improvements have typically been demonstrated on a very narrow set of tasks with no rigorous testing on different linguistic subfields (Moussa et al., 2025; Vattikonda et al., 2025; Freteault et al., 2025; Negi et al., 2025), making it unclear whether the benefits of brain-tuning generalize across linguistic competencies.

In this work, we provide a direct and controlled test of the value of brain data for text-based language model training. To this end, we compare three types of models: Brain-Tuned LMs, Stimulus-Tuned LMs (Merlin and Toneva, 2024) (i.e., models trained on text corresponding to the linguistic stimulus underlying the corresponding brain data used by the Brain-Tuned models) and Jointly-Tuned LMs (i.e., models trained on brain recordings and text corresponding to the linguistic stimulus)(Merlin and Toneva, 2026). This design isolates the contribution of brain data itself, enabling a direct comparison between learning from text alone, learning from brain signals and learning from text augmented with neural signals. We then evaluate Brain-Tuned, Stimulus-Tuned and Jointly-Tuned models across a broad suite of 200 downstream

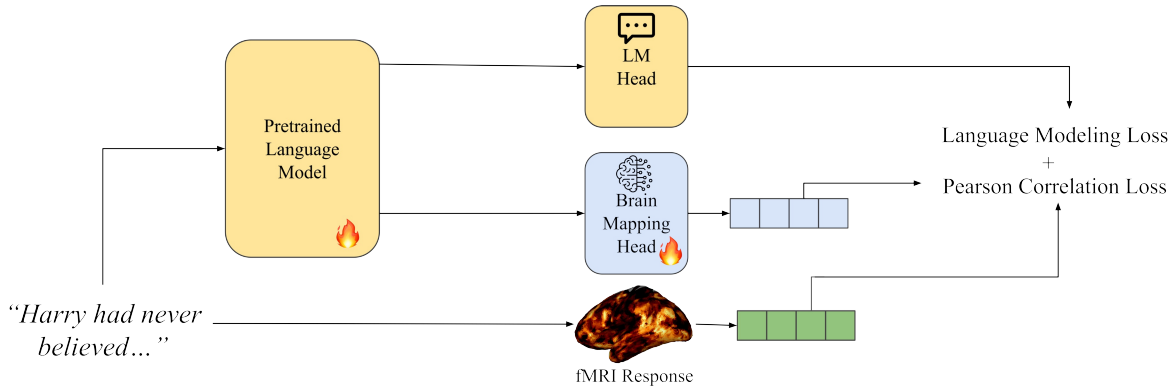

Figure 1: Overview of the proposed approach. Our results are based on fine-tuning LMs with two objectives: language modeling ability and alignment with brain recordings. The two objectives are used simultaneously for the Jointly-Tuned Model or individually for Brain-Tuned and Stimulus-Tuned models.

language tasks across syntax, semantics, discourse, morphology, and reasoning (Waldis et al., 2024). The comparison between Jointly-Tuned and pretrained models establishes the overall utility of incorporating brain-related supervision. The comparison between Brain-Tuned and Stimulus-Tuned models isolates the unique contribution of neural signals beyond additional exposure to stimulus text. Together, these contrasts allow us to disentangle improvements due to extra textual training from those attributable specifically to neural data.

Our results show that Jointly-Tuned LMs consistently outperform pretrained baselines, highlighting the importance of brain-related information as a supervisory signal for LMs. Moreover, by comparing Brain-Tuned models with Stimulus-Tuned ones, we show that the former outperform the latter. These gains are observed across multiple types of downstream tasks, demonstrating that brain data provide a rich and complementary supervision signal that cannot be reduced to additional exposure to text. Together, our findings establish that neural recordings offer more than a proxy for linguistic input: they encode information that improves generalization in language models.

Our main contributions are:

1. We conduct the first comprehensive evaluation of Jointly-Tuned, Brain-Tuned and Stimulus-Tuned text-based language models on linguistic competence, systematically analyzing performance across over 200 probing tasks spanning syntax, semantics, discourse, reasoning, and morphology using the Holmes benchmark.

2. We show that brain recordings provide complementary information beyond textual stimuli: Jointly-Tuned models outperform pretrained models in every linguistic subfield, as do Brain-Tuned models compared to matched Stimulus-Tuned controls[1].

2 Related works

Brain-language model alignment. A substantial literature documents alignment between pretrained language models and human brain activity during natural language comprehension: model representations can predict fMRI or MEG responses above chance when humans and models process the same stimuli (Wehbe et al., 2014b; Jain and Huth, 2018; Toneva and Wehbe, 2019; Abdou et al., 2021; Schrimpf et al., 2021; Hosseini et al., 2024). Beyond establishing that alignment exists, several studies investigate its sources, linking representational properties and architectural choices to neural responses (Goldstein et al., 2022; Toneva et al., 2022a; Oota et al., 2024a,b; Caucheteux et al., 2021; Reddy and Wehbe, 2021; Toneva et al., 2022b; Kauf et al., 2023; Gauthier and Levy, 2019; Aw and Toneva, 2023; Merlin and Toneva, 2024). These works establish the existence of brain-model alignment but leave open whether this alignment is functionally relevant for language processing capabilities. Recent efforts have begun to address this by explicitly manipulating the link between brain and model representations. For instance, Merlin and Toneva (2026) investigated the functional implications of alignment by introducing "Brain-Misaligned" models—LLMs intentionally trained

[1]Code is publicly available at https://github.com/bridge-ai-neuro/lm-brain-tuning

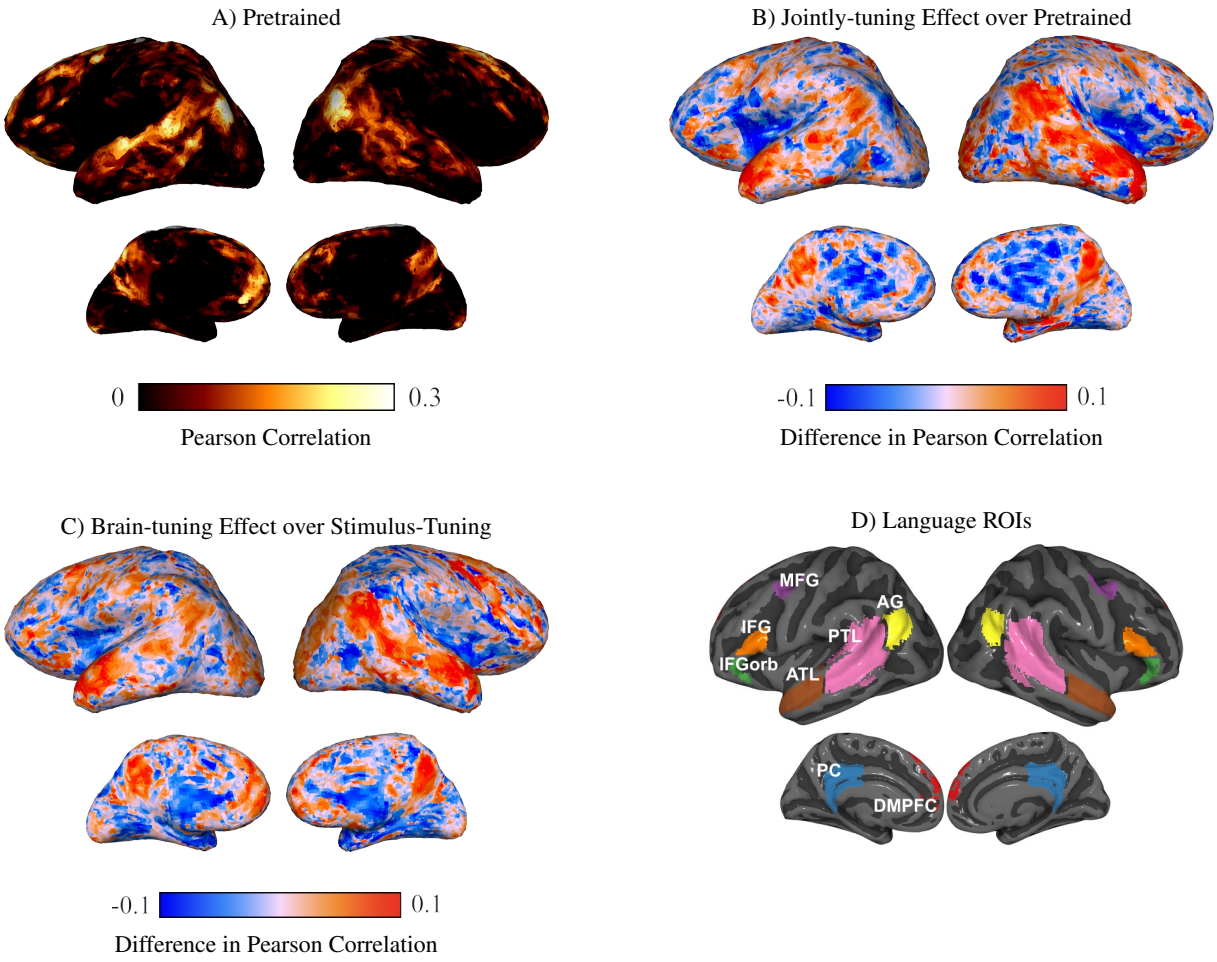

Figure 2: Brain alignment of the BERT pretrained model (A) for one participant on the Harry Potter dataset (see Appendix F for all participants), the effect of Jointly-Tuning the model over the pretrained(B), and the effect of Brain-tuning the model over Stimulus-Tuning(C). The Jointly-Tuned model and Brain-Tuned models exhibits substantially stronger alignment, particularly in language regions (C, D).

to predict brain activity poorly while maintaining high language modeling performance. Their results reveal that brain misalignment substantially impairs downstream performance across diverse linguistic domains, and that jointly-tuning with brain data and corresponding stimuli can induce improvement in linguistic performance, suggesting that this alignment is critical for the models' robust linguistic competence. However, their experimental design leaves the specific contribution of brain data unclear. Since their models are fine-tuned using a combined loss that incorporates both neural recordings and the underlying textual stimuli simultaneously, the resulting performance gains remain entangled.

Brain-tuning of language models. In addition, recent work built on this alignment to explore whether brain signals can actively improve language models through fine-tuning. Schwartz et al. (2019) pioneered this direction, demonstrating that fine-tuning BERT with fMRI and MEG recordings improved both brain encoding accuracy and performance on selected NLP tasks. More recently, Moussa et al. (2025) extended brain-tuning to speech models, showing improved semantic alignment and consistent gains on semantic downstream tasks. Negi et al. (2025) further explored multilingual brain-tuning effects. However, these works have primarily focused on speech models or had limited downstream evaluation, leaving open whether brain-tuning provides broad benefits for linguistic competence in language models across diverse linguistic subfields. Moreover, prior text-based studies have mostly compared Brain-Tuned models against their pretrained versions, yielding mixed results and making it difficult to attribute the results to the neural signals or to seeing more text during finetuning.

Probing linguistic competence. To address this gap, we draw on complementary work that systematically evaluates what language models know about language structure and meaning, using controlled evaluation frameworks to measure model performance across syntactic, semantic, morphological, discourse, and reasoning tasks (Amouyal et al., 2024; Blevins et al., 2023). Key benchmarks include BLiMP for grammatical knowledge through minimal pairs (Warstadt et al., 2020), GLUE and SuperGLUE for diverse natural language understanding tasks (Wang et al., 2018, 2019), and Holmes, which organizes over 200 probing datasets by linguistic subdomain for comprehensive competence evaluation (Waldis et al., 2024).

Our study complements this literature by focusing on language models and by evaluating the impact of jointly-tuning and brain-tuning on a broad, linguistically structured benchmark rather than on a small set of tasks, connecting alignment-driven training to measurable changes in linguistic competence and exploring what finetuning on brain data only can add to LMs in addition to textual data. Recent work also links brain alignment primarily to *formal* linguistic competence: syntax, morphology, and word-level lexical semantics (rather than *functional* competence such as pragmatic reasoning and real-world language use) (Mahowald et al., 2024; AlKhamissi et al., 2025). The phenomena where our Brain-Tuned and Jointly-Tuned models show the largest gains fall within this formal-competence domain, in line with these previous works.

3 Methodology

3.1 Model Selection and Architecture

We use two widely-studied transformer architectures: BERT (Devlin et al., 2019) and GPT-2 (Radford et al., 2019). Specifically, we work with `bert-base-cased` and `gpt2-small` from the Transformers library (Wolf et al., 2020). These models have established track records both in downstream NLP performance and in neuroscience studies examining brain-model alignment (Toneva and Wehbe, 2019; Caucheteux et al., 2021; Oota et al., 2024b).

3.2 Brain Data

Our experiments are based on two publicly available fMRI datasets. One dataset is provided by Wehbe et al. (2014a). This dataset captures neural responses from eight participants during word-by-word reading of Chapter 9 from Harry Potter and the Sorcerer's Stone (Rowling et al., 1998). The experimental protocol presented each word for 500ms with fMRI volumes acquired every 2 seconds (TR=2s), yielding 1211 brain images per participant across four experimental runs. The second dataset, provided by Deniz et al. 2019, consists of brain recordings from six participants who read and listened to 11 stories from The Moth Radio Hour. For both reading and listening, the dataset includes 4028 fMRI images. In our experiments we use the reading data, in which each word was presented for the same duration as in the audio recording. We selected this dataset for its combination of naturalistic stimuli, large per-participant data volume, and established use in brain-language model alignment research.

3.3 Brain-Tuning Framework

To isolate the contribution of brain-related information to linguistic competence, we implement a controlled experimental design comparing three fine-tuning approaches. Our Brain-Tuned models are optimized using fMRI recordings, Stimulus-Tuned baseline models are trained exclusively on the corresponding textual stimuli and Jointly-Tuned models are trained on both textual stimuli and fMRI recording. This design allows us to disentangle the specific value of neural signals from general effects of continued training with stimuli.

3.3.1 Brain-Tuned Models

Our Brain-Tuned models incorporate brain-relevant information through fine-tuning. We add a specialized prediction head (the *brain mapping head* in Figure 1) that maps model representations to brain activity patterns. During fine-tuning, the model optimizes the brain prediction objective using fMRI data of a single participant from the fMRI dataset discussed in Section 3.2.

For the brain prediction component, we select voxels with noise ceiling values exceeding 0.05 (detailed in Appendix D) from established language-processing regions (Fedorenko et al., 2010; Fedorenko and Thompson-Schill, 2014; Binder et al., 2009) and made available from Wehbe et al. (2014a) and Deniz et al. (2019). The brain prediction loss is computed as the mean Pearson correlation between predicted and actual voxel responses across each batch. Implementation details for the prediction head are provided in Appendix B.

3.3.2 Stimulus-Tuned Models

Our Stimulus-Tuned control models provide a matched baseline for the Brain-Tuned models by training on corresponding textual content. These models undergo fine-tuning using only a language modeling objective, applied to the same text sequences that participants read during fMRI scanning. We use the same language modeling loss used to train the pretrained models. This control ensures that any performance differences reflect the contribution of brain signals rather than exposure to additional textual data.

3.3.3 Jointly-Tuned Models

This model was finetuned using a combination of two prediction heads, one related to language modeling as discussed in 3.3.2 and a brain prediction head as described in section 3.3.1. The total loss is defined as:

$$\mathcal{L} = \omega_{lm} * \mathcal{L}_{lm} + \omega_{ba} * \mathcal{L}_{ba}$$

where $\mathcal{L}_{lm}$ is the language modeling loss, $\mathcal{L}_{ba}$ is the brain-alignment loss, and ω_{lm} and ω_{ba} are weighting factors to balance the two objectives.

3.3.4 Training Configuration

Training samples consist of word sequences spanning 5 TRs. We partition the data into four consecutive segments to enable cross-validation across different story sections.

We fine-tune models using LoRA (Hu et al., 2022) for 5 epochs. In preliminary tests, LoRA yielded more stable training dynamics than full-parameter fine-tuning (training details are reported in Appendix C). For the Brain-Tuned model for each participant and data split (8 participants $\times$ 4 subsets for Harry Potter dataset and 6 participants $\times$ 4 subsets for the Moth Radio Hour dataset), we select the checkpoint that maximizes brain alignment on a validation set (20% of the full training data). After selection, the model is retrained on the full training data and the chosen checkpoint is retained for evaluation.

3.4 Evaluation

We evaluate our models in two main ways to assess both the effectiveness of brain-tuning and its impact on linguistic capabilities. All evaluations use held-out data that was not seen during fine-tuning.

Brain alignment evaluation. To quantify how well our models capture brain activity patterns, we employ linear encoding models that map from model representations to voxel responses. Following established methods (Toneva and Wehbe, 2019; Schrimpf et al., 2021), we extract representations from the final transformer layer and train ridge-regularized linear mappings to predict fMRI responses. The ridge regularization parameter is selected through nested cross-validation to prevent overfitting.

For each participant and held-out run, we train separate encoding models and compute Pearson correlations between predicted and actual voxel responses. Brain alignment for model q and voxel v_j is defined as:

$$\texttt{brain alignment}(q, v_j) = \texttt{corr}(\hat{y_j}, y_j)$$

where $\hat{y_j}$ represents the predicted response and y_j the ground-truth fMRI signal.

Linguistic competence evaluation. To measure downstream linguistic capabilities, we employ the Holmes benchmark (Waldis et al., 2024), which encompasses over 200 probing tasks across five major linguistic domains: syntax, semantics, morphology, discourse, and reasoning (see E for details). Each task involves training classifiers on frozen model representations to predict linguistic properties or judgments.

We run each task with 6 different random seeds to account for variability in probe initialization and data ordering. Statistical significance is assessed using two-sample t-tests comparing performance distributions between models. We assign a "win" to a model only when it significantly outperforms others ($p < 0.05$). This approach yields a filtered win-rate that prioritizes robustness, ensuring that reported improvements reflect consistent advantages rather than stochastic noise.

Since we train multiple model pairs across different cross-validation folds, we aggregate win rates across folds for each participant and dataset. This yields a win score indicating how consistently one training approach outperforms the others across different data splits and random seeds.

4 Results

4.1 Effects on Brain Alignment

Figure 2A–B illustrates brain alignment for the pretrained BERT model and the difference with the BERT-based Jointly-Tuned, trained on the Harry Potter data for one representative participant. Alignment is quantified as the Pearson cor-

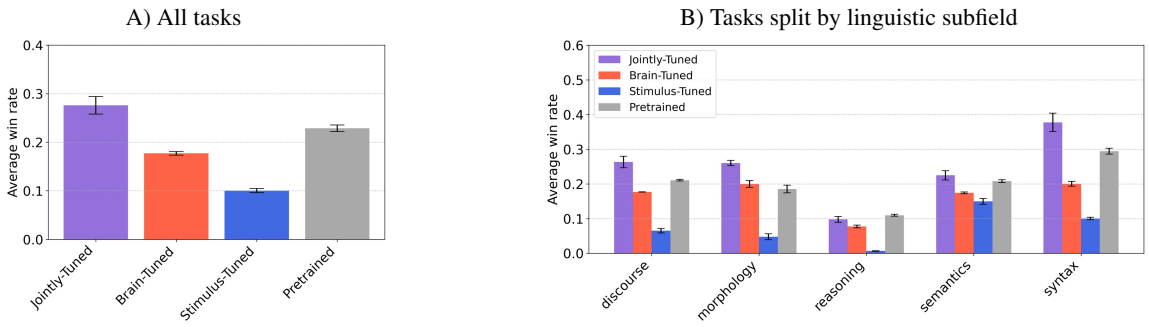

Figure 3: Average win rate and standard error across model and dataset combinations for Jointly-Tuned, Brain-Tuned, and Stimulus-Tuned models, shown across all tasks (Left) and linguistic subfields (Right). The win rate indicates how often each model outperforms others across tasks and participants. The Jointly-Tuned model consistently outperforms all other models across each model-dataset combination; in particular, it outperforms Brain-Tuned and Stimulus-Tuned models ($p < 0.05$, Wilcoxon signed-rank test). The Brain-Tuned models significantly outperforms the Stimulus-Tuned models (Left), suggesting that enforcing brain alignment positively influences linguistic competence. The Jointly-Tuned model shows a higher win rate in all linguistic subfields (Right) and is particularly strong in syntax, morphology, and discourse ($p < 0.05$, Wilcoxon signed-rank test with Holm-Bonferroni correction), suggesting that brain alignment particularly affects linguistic competence in these subfields.

relation between predicted and observed voxel responses. Figure 2A reports voxel-wise correlations for the pretrained models, Figure 2B displays the difference between Jointly-Tuned and pretrained models and Figure 2C shows the difference between Brain-Tuned and Stimulus-Tuned models. Language-related regions of interest, used to guide our analyses, are shown in Figure 2D. Additional participants, models, and datasets are reported in Appendix F, G and quantified in Figure 9, 14, 19, 24.

Pretrained Model. Pretrained models show robust alignment across the cortex, with the strongest effects concentrated in language-sensitive areas such as the left inferior frontal gyrus and superior temporal regions, in agreement with prior work (Fedorenko et al., 2010; Fedorenko and Thompson-Schill, 2014; Binder et al., 2009).

Jointly-Tuning Effect. The Jointly-Tuned models show improvement across the cortex, with the strongest effects concentrated in language-sensitive areas such as the left inferior frontal gyrus and superior temporal regions. These findings confirm that neural and textual supervision allows the model to capture features that reflect the brain activity observed during the fine-tuning process.

Brain-Tuning Effect over the Stimulus-Tuned model. Figure 2C illustrates the voxel-wise contrast between Brain-Tuned and Stimulus-Tuned models for a representative subject. The most pronounced differences are localized within language regions (Figure 2D).

Effects on other subjects and models. Across participants, BERT-based Brain-Tuned models generally outperform both the Stimulus-Tuned and the Pretrained baselines (Figure 9, 14, 24, 19). The Jointly-Tuned models show gains over the pretrained models, particularly regarding the alignment in the Moth Radio Hour dataset; however, for the Harry Potter dataset, mixed results are observable, likely due to the smaller size of the dataset.

4.2 Effects on Linguistic Competence

Figures 3A, 3B, and 4 summarize the comparison between Brain-Tuned, Stimulus-Tuned, Jointly-Tuned models on the Holmes benchmark, averaged across models and datasets. Figure 3A reports the aggregated effect across all tasks, Figure 3B breaks down performance by linguistic subfield, and Figure 4 highlights specific linguistic phenomena. Results for particular models and datasets are provided in Appendix (Section F G, H, I).

Overall linguistic competence. As shown in Figure 3A, Jointly-Tuned models achieve a higher average win rate across the full benchmark compared to the pretrained model. This result is consistent across models and datasets. Such findings suggest that the integration of brain-related signals together with stimuli during training elicits superior performance in linguistic tasks compared to pretrained models. Brain-Tuned and Stimulus-Tuned models are significantly worse than the pretrained model,

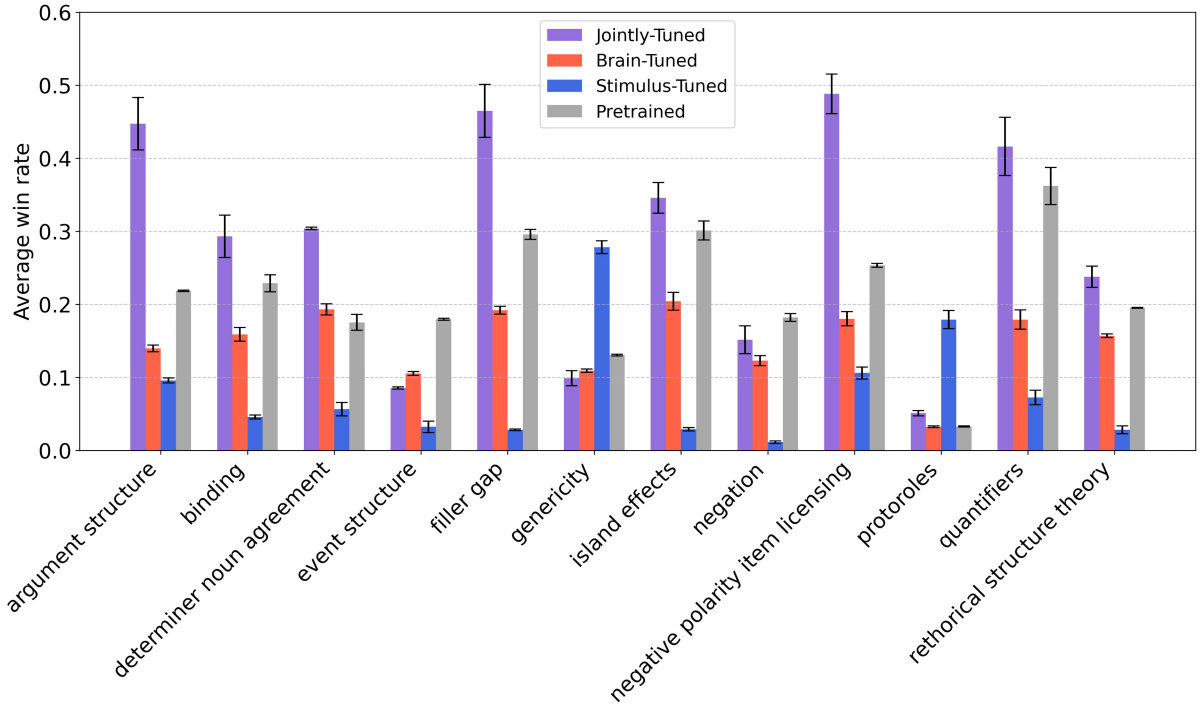

Figure 4: Average win rate and standard error across model and dataset combinations, across various linguistic phenomena, for the Brain-Tuned, Stimulus-Tuned models and Jointly-Tuned models. Each bar represents the average win rate for a specific linguistic phenomenon, with error bars indicating standard error. Jointly-Tuned models outperform pretrained ones in the majority of linguistic phenomena. In the majority of linguistic phenomena the Brain-Tuned models show significant improvement (asterisks indicate statistically significant differences $p < 0.05$, paired t-test with Holm-Bonferroni correction). Some concrete examples of the linguistic tasks are provided in the Table 2.

suggesting that fine-tuning on a specific task alone does not help in improving overall linguistic competence. This performance degradation is likely attributable to a combination of catastrophic forgetting and the extremely limited size of the stimulus datasets. However, Brain-Tuned models show better performance than Stimulus-Tuned models, demonstrating that brain data has an added value with respect to textual data.

Subfield-level performance. The breakdown by linguistic subfield in Figure 3B shows consistent advantages for Jointly-Tuned models across every category with respect to pretrained model. The differences are especially strong for syntax, morphology, and discourse. This pattern suggests that fine-tuning with brain data and corresponding stimuli particularly reinforces abilities related to these linguistic subfields. Brain-Tuned models show better performance than Stimulus-Tuned models in every linguistic subfield, in particular for syntax, reasoning, morphology, and discourse. This demonstrates that models tuned with brain data can outperform models tuned on textual data. Brain-Tuned and Stimulus-Tuned models are generally worse than the pretrained ones in the majority of linguistic subfields, with the exception of morphology, where Brain-Tuned models outperform the pretrained baseline. This suggests that while overall performance is lower, using brain data during fine-tuning helps improve performance for this particular linguistic subfield.

Phenomenon-level performance. To obtain a finer-grained view, we further analyzed linguistic phenomena represented by more than five datasets. As illustrated in Figure 4, Jointly-Tuned models tend to outperform pretrained ones across most categories. The largest gains appear in tasks involving *argument structure*, *determiner noun agreement*, *filler–gap*, *negative polarity*, and *rhetorical structure*. Brain-Tuned models outperform Stimulus-Tuned models on the majority of linguistic phenomena. These results show that brain data used during fine-tuning can incorporate important information for performing linguistic tasks that text data alone cannot provide. Brain-Tuned and Stimulus-Tuned models are generally worse than the pre-

trained ones, except for *determiner noun agreement*, *genericity* and *protoroles*. Examples of such phenomena are listed in Table 2.

5 Discussion

In this work, we investigated the impact of brain tuning on language models. We compared three models: Jointly-Tuned models, that incorporate fMRI recordings and the corresponding textual stimuli during the training process, Brain-Tuned models, which incorporate only fMRI recordings, and Stimulus-Tuned models, trained solely on the corresponding textual stimuli.

Our results show that integrating brain signals and textual stimuli during finetuning leads to consistent improvements in downstream linguistic tasks. These gains are observable in every linguistic subfield (Figure 3), suggesting that brain data and corresponding textual stimuli provide complementary information that enhances the model's linguistic representations.

To better understand which of the two training signals guide the improvements in linguistic competence we compared Brain-Tuned models with Stimulus-Tuned models. The higher performance of Brain-Tuned models over Stimulus-Tuned suggests that neural signals provide unique relevant information that is not available in raw text.

Brain-Tuned and Stimulus-Tuned models are generally worse than pretrained ones, with the Brain-Tuned model being better than the pretrained baseline only for morphology tasks. These results suggest that single-task fine-tuning might not be enough to improve linguistic performance. Furthermore, the small size of the dataset and excessive specialization during fine-tuning could cause the model to diverge from representations that are useful for generalizability across different tasks.

Overall, our findings support the hypothesis that brain recordings capture aspects of language processing not fully represented in textual data. By leveraging this complementary signal, language models can achieve improved alignment with human cognition and enhanced linguistic competence. Brain-tuning serves as a methodological tool to integrate biological inductive biases into model representations, thereby contributing to the development of LLMs as "model organisms" of the human brain. This approach enables an alignment between LLMs and the human brain that goes beyond a shared capacity for textual comprehension, specifically targeting representational similarity. In this light, brain data provides a high-dimensional proxy for human language processing, allowing for an evaluation of how enforcing representational similarity with the brain impacts linguistic competence. These results open new avenues for incorporating neural data into language model training, suggesting that even small amounts of brain recordings can meaningfully enrich model representations.

Limitations. Our study has three main limitations. Firstly, the benchmark used to assess linguistic competence, while extensive, is not exhaustive. There are many additional datasets available that could be included in future evaluations. Moreover, some linguistic subfields (e.g., discourse) and specific linguistic phenomena are represented by only a few datasets. As a result, the observed behavior of the Brain-Tuned and Stimulus-Tuned models may be influenced by the limited coverage and distribution of tasks in certain categories. Secondly, our results are based on limited fMRI datasets. Although these datasets are among the largest available in terms of data per participant and have been widely studied in previous work, the findings may still be specific to its characteristics. Thirdly, we designed our experiments using cross-validation, testing on held-out data and across multiple participants to improve generalizability. However, results might differ with different text genres or other types of linguistic stimuli. Expanding to more datasets, languages, or cognitive tasks could be an important next step. Nevertheless, our results are informative in demonstrating the effectiveness of our methodology and in highlighting the importance of the emergent brain alignment ability of LMs.

6 Conclusion

In this work, we presented an investigation into the functional benefits of incorporating human brain signals into language model training. By introducing an experimental framework that compares Jointly-Tuned models against pretrained models and Brain-Tuned against Stimulus-Tuned models, we successfully isolated the unique contribution of neural data from the effects of simple exposure to additional linguistic material.

Our evaluation across the Holmes benchmark—covering over 200 datasets—demonstrates that jointly-tuning enhances linguistic competence of language models. The consistent performance gains observed in particular for syntax, morphol-

ogy and discourse, suggest that fMRI recordings and corresponding stimuli provide a rich, complementary supervisory signal. By isolating the contribution of brain-tuning alone, we show that it offers a more informative signal than pure fine-tuning with textual stimuli. Moreover, compared to pretrained performance, only Jointly-Tuned models are able to consistently outperform the baseline, indicating that brain and stimulus data complement each other and contribute to better linguistic competence. While previous studies have largely focused on observing the emergence of brain-model alignment, our results provide empirical evidence that this alignment is functionally relevant: models that are explicitly finetuned to align with brain recordings develop representations that incorporate important linguistic information.

Our methodology provides a general framework for assessing the utility of emergent properties in language models by disentangling their effects from those of linguistic exposure. This comparative paradigm allows for a rigorous quantification of the "information gain" provided by non-textual signals. While we focused on fMRI data in the context of natural language, this framework is inherently flexible and can be extended to other modalities, such as vision or audio, as well as alternative brain imaging techniques including EEG, MEG, or ECoG.

Acknowledgements

This work was funded in part by the German Research Foundation (DFG) - DFG Research Unit FOR 5368 and by the CS@max planck graduate center.

References

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

Badr AlKhamissi, Greta Tuckute, Antoine Bosselut, and Martin Schrimpf. 2025. From language to cognition: How llms outgrow the human language network. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Samuel Amouyal, Aya Meltzer-Asscher, and Jonathan Berant. 2024. Large language models for psycholinguistic plausibility pretesting. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 166–181, St. Julian's, Malta. Association for Computational Linguistics.

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

Jeffrey R Binder, Rutvik H Desai, William W Graves, and Lisa L Conant. 2009. Where is the semantic system? a critical review and meta-analysis of 120 functional neuroimaging studies. *Cerebral cortex*, 19(12):2767–2796.

Terra Blevins, Hila Gonen, and Luke Zettlemoyer. 2023. Prompting language models for linguistic structure. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6649–6663, Toronto, Canada. Association for Computational Linguistics.

Charlotte Caucheteux, Alexandre Gramfort, and Jean-Remi King. 2021. Disentangling syntax and semantics in the brain with deep networks. In *International conference on machine learning*, pages 1336–1348. PMLR.

Fatma Deniz, Anwar O Nunez-Elizalde, Alexander G Huth, and Jack L Gallant. 2019. The representation of semantic information across human cerebral cortex during listening versus reading is invariant to stimulus modality. *Journal of Neuroscience*, 39(39):7722–7736.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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)*, pages 4171–4186.

E. Fedorenko, P.-J. Hsieh, A. Nieto-Castanon, S. Whitfield-Gabrieli, and N. Kanwisher. 2010. New method for fMRI investigations of language: Defining ROIs functionally in individual subjects. *Journal of Neurophysiology*, 104(2):1177–1194.

Evelina Fedorenko and Sharon L Thompson-Schill. 2014. Reworking the language network. *Trends in cognitive sciences*, 18(3):120–126.

Maëlle Freteault, Maximilien Le Clei, Loic Tetrel, Lune Bellec, and Nicolas Farrugia. 2025. Alignment of auditory artificial networks with massive individual fmri brain data leads to generalisable improvements in brain encoding and downstream tasks. *Imaging Neuroscience*, 3.

Jon Gauthier and Roger Levy. 2019. Linking artificial and human neural representations of language. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 529–539, Hong Kong, China. Association for Computational Linguistics.

Ariel Goldstein, Zaid Zada, Eliav Buchnik, Mariano Schain, Amy Price, Bobbi Aubrey, Samuel A Nastase, Amir Feder, Dotan Emanuel, Alon Cohen, and 1 others. 2022. Shared computational principles for language processing in humans and deep language models. *Nature neuroscience*, 25(3):369–380.

Eghbal A. Hosseini, Martin Schrimpf, Yian Zhang, Samuel Bowman, Noga Zaslavsky, and Evelina Fedorenko. 2024. Artificial Neural Network Language Models Predict Human Brain Responses to Language Even After a Developmentally Realistic Amount of Training. *Neurobiology of Language*, 5(1):43–63.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

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

Shailee Jain and Alexander Huth. 2018. Incorporating context into language encoding models for fMRI. In *Advances in Neural Information Processing Systems*, pages 6628–6637.

Carina Kauf, Greta Tuckute, Roger Levy, Jacob Andreas, and Evelina Fedorenko. 2023. Lexical semantic content, not syntactic structure, is the main contributor to ann-brain similarity of fmri responses in the language network. *bioRxiv*.

Kyle Mahowald, Anna A. Ivanova, Idan A. Blank, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. 2024. Dissociating language and thought in large language models. *Trends in Cognitive Sciences*, 28(6):517–540.

Gabriele Merlin and Mariya Toneva. 2024. Language models and brains align due to more than next-word prediction and word-level information. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18431–18454, Miami, Florida, USA. Association for Computational Linguistics.

Gabriele Merlin and Mariya Toneva. 2026. When language models lose their mind: The consequences of brain misalignment. In *The Fourteenth International Conference on Learning Representations*.

Omer Moussa, Dietrich Klakow, and Mariya Toneva. 2025. Improving semantic understanding in speech language models via brain-tuning. *ICLR*.

Omer Moussa and Mariya Toneva. 2025. Brain-tuning improves generalizability and efficiency of brain alignment in speech models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Anuja Negi, Subba Reddy Oota, Anwar O Nunez-Elizalde, Manish Gupta, and Fatma Deniz. 2025. Brain-informed fine-tuning for improved multilingual understanding in language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Shinji Nishimoto, An T Vu, Thomas Naselaris, Yuval Benjamini, Bin Yu, and Jack L Gallant. 2011. Reconstructing visual experiences from brain activity evoked by natural movies. *Current biology*, 21(19):1641–1646.

Subba Reddy Oota, Emin Çelik, Fatma Deniz, and Mariya Toneva. 2024a. 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)*, pages 8503–8528, Bangkok, Thailand. Association for Computational Linguistics.

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

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, and 1 others. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Aniketh Janardhan Reddy and Leila Wehbe. 2021. Can fmri reveal the representation of syntactic structure in the brain? *Advances in Neural Information Processing Systems*, 34:9843–9856.

J.K. Rowling, M. GrandPre, M. GrandPré, T. Taylor, Arthur A. Levine Books, and Scholastic Inc. 1998. *Harry Potter and the Sorcerer's Stone*. Harry Potter. A.A. Levine Books.

Martin Schrimpf, Idan Asher Blank, Greta Tuckute, Carina Kauf, Eghbal A Hosseini, Nancy Kanwisher, Joshua B Tenenbaum, and Evelina Fedorenko. 2021. The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, 118(45).

Dan Schwartz, Mariya Toneva, and Leila Wehbe. 2019. Inducing brain-relevant bias in natural language processing models. *Advances in neural information processing systems*, 32.

Mariya Toneva, Tom M Mitchell, and Leila Wehbe. 2022a. Combining computational controls with natural text reveals aspects of meaning composition. *Nature Computational Science*, 2(11):745–757.

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

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

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

Andreas Waldis, Yotam Perlitz, Leshem Choshen, Yufang Hou, and Iryna Gurevych. 2024. Holmes a benchmark to assess the linguistic competence of language models. *Transactions of the Association for Computational Linguistics*, 12:1616–1647.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, 32.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R Bowman. 2020. Blimp: The benchmark of linguistic minimal pairs for english. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Leila Wehbe, Brian Murphy, Partha Talukdar, Alona Fyshe, Aaditya Ramdas, and Tom Mitchell. 2014a. Simultaneously uncovering the patterns of brain regions involved in different story reading subprocesses. *PloS one*, 9(11):e112575.

Leila Wehbe, Ashish Vaswani, Kevin Knight, and Tom Mitchell. 2014b. 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*, pages 233–243.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 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*, pages 38–45, Online. Association for Computational Linguistics.

A Appendix

B Brain Mapping Head

To predict fMRI activity for each TR, we use a ridge-regularized linear mapping from model representations to voxel responses. We restrict the prediction to voxels with a noise ceiling above 0.05 in language-related regions of interest (Figure 2D). The mapping is trained with cross-validation and evaluated on held-out data, with the ridge parameter chosen via nested cross-validation.

For each participant, we train four mappings, each using three of the four fMRI subsets for training and the remaining one for testing. Model representations are obtained by averaging token embeddings within a TR. The input features are constructed by concatenating embeddings from the current TR with those from the previous five TRs, in order to account for the delay of the hemodynamic response. Since fMRI signals reflect neural activity indirectly and peak around 6 seconds after stimulus onset, including preceding TRs allows the model to estimate voxel-specific hemodynamic response functions (HRFs) in a data-driven manner (Nishimoto et al., 2011; Wehbe et al., 2014a; Huth et al., 2016).

C Implementation Details and Evaluation

Jointly-Tuned models are fine-tuned using a learning rate of 5×10^{-4} and Low-Rank Adaptation (LoRA) (Hu et al., 2022) with a rank $r = 8$ and the language modeling loss weight $\omega_{lm} = 0.1$ and $\omega_{ba} = 10$ values chosen based on the relative magnitude of the losses prior to fine-tuning. For Brain-Tuned and Stimulus-Tuned models we use a learning rate of 5×10^{-5} and Low-Rank Adaptation (LoRA) (Hu et al., 2022) with a rank $r = 4$. These parameters are selected by using a validation set on a subset of subjects. For Brain-Tuned and Jointly-Tuned models we use the Pearson correlation (ρ) within each batch, as objective function. For a batch of size B, the correlation is defined as:

$$\rho = \frac{\sum_{i=1}^{B}(y_i - \mu_{\mathbf{y}})(\hat{y}_i - \mu_{\hat{\mathbf{y}}})}{\sqrt{\sum_{i=1}^{B}(y_i - \mu_{\mathbf{y}})^2 \sum_{i=1}^{B}(\hat{y}_i - \mu_{\hat{\mathbf{y}}})^2}} \tag{1}$$

where $\mu_{\mathbf{y}}$ and $\mu_{\hat{\mathbf{y}}}$ represent the batch-wise means for the ground truth and predicted signals, respectively.

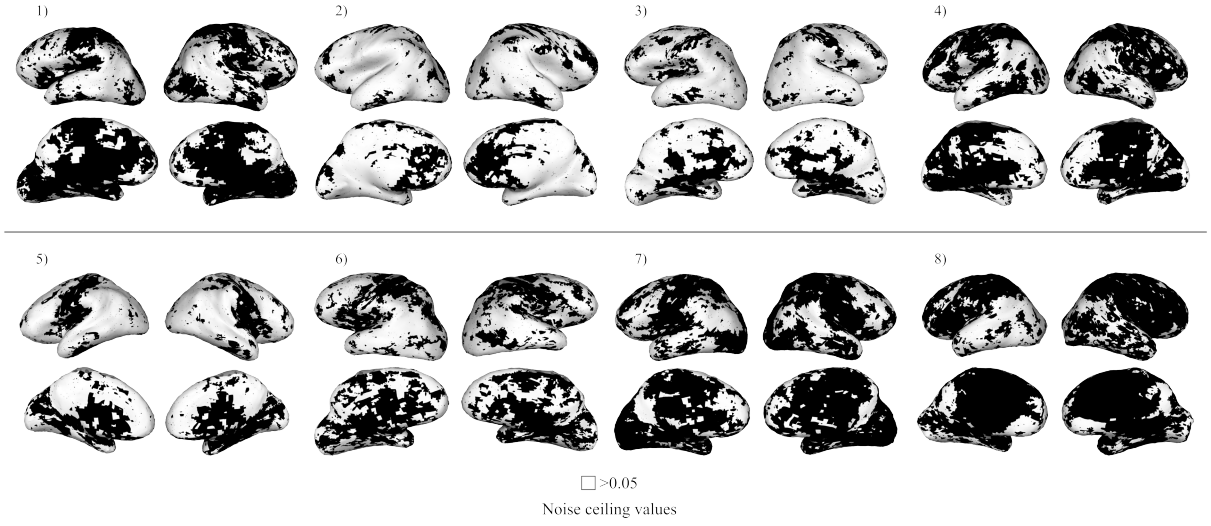

Figure 5: Voxel-wise noise ceiling estimates for participants in the Harry Potter dataset (Wehbe et al., 2014a). For each participant, we retained only voxels with a noise ceiling above 0.05.

D Noise Ceiling estimation

To evaluate fMRI signal quality, we estimated voxel-wise noise ceilings, which measure the fraction of variance explainable by an ideal model. Following Schrimpf et al. (2021), this is done by predicting a participant's fMRI responses using linear models trained on data from another participant. Because fMRI

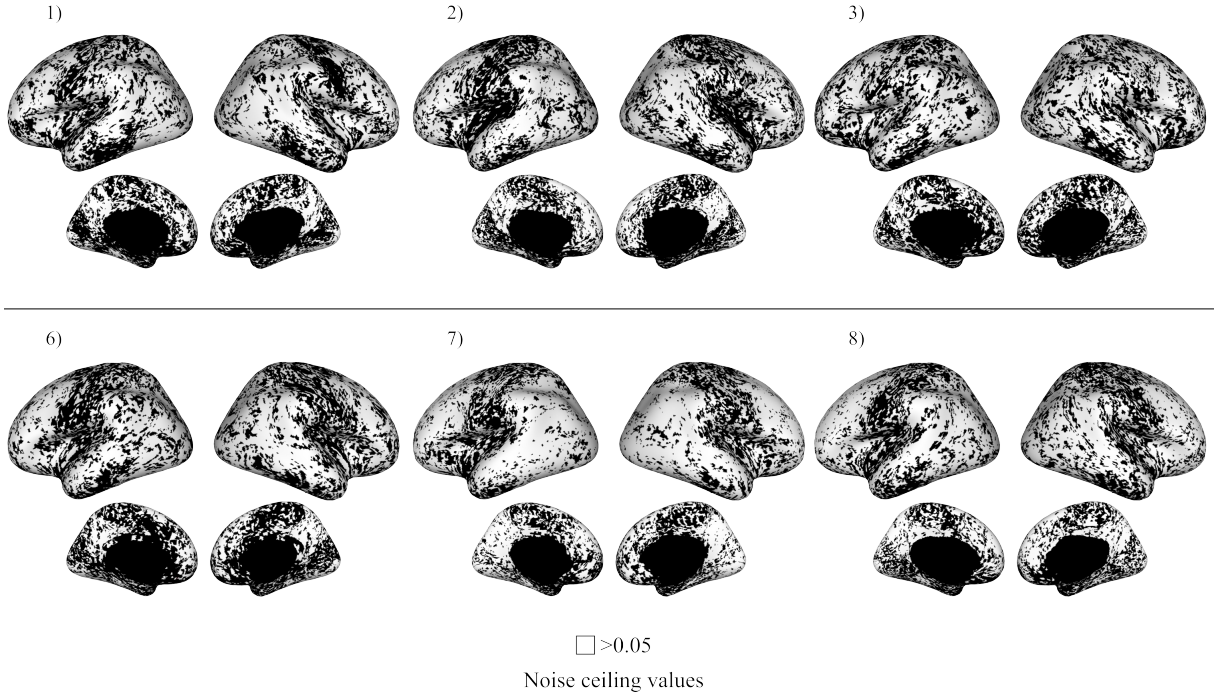

Figure 6: Voxel-wise estimated noise ceiling values for participants included in Moth Radio Hour dataset (Deniz et al., 2019). For each participant, we retained only voxels with a noise ceiling above 0.05.

Table 1: Examples for linguistic subfields from (Waldis et al., 2024). The relevant part of the example for the specific label is underlined.

Type	Phenomena	Example	Label
Morphology	Subject-Verb Agreement	*And then, the cucumber was hurled into the air.*	`Correct`
		And then, the cucumber were hurled into the air.	`Wrong`
Syntax	Part-of-Speech	*And then, the cucumber was hurled into the air.*	`NN (Noun Singular)`
Semantics	Semantic Roles	*And then, the cucumber was hurled into the air.*	`Direction`
Reasoning	Negation	*And then, the cucumber was hurled into the air.*	`No Negation`
Discourse	Node Type in Rhetorical Tree	*And then, the cucumber was hurled into the air.*	`Satellite`

data are inherently noisy, estimating the noise ceiling provides a principled upper bound on predictive accuracy.

E Linguistic competence benchmark

We assessed linguistic competence using classifier-based probing on the Holmes benchmark (Waldis et al., 2024), which aggregates over 200 datasets from diverse sources. Holmes provides broad coverage of linguistic knowledge, with tasks spanning syntax, morphology, semantics, reasoning, and discourse. A detailed mapping of linguistic phenomena to subfields is given in Table 3. The benchmark includes diverse classification paradigms depending on the target phenomenon, including token/span-level categorization (e.g., classifying the part-of-speech tag of a specific word) and sequence-level classification (e.g., predicting binary acceptability judgments for a full sentence).

For efficiency, we used the `flash-holmes` version of the benchmark, which substantially reduces computational cost while preserving the reliability of competence estimates (see (Waldis et al., 2024) for details). Representative tasks for each phenomenon are reported in Table 2.

Table 2: Examples for selected linguistc phenomena from (Waldis et al., 2024). The asterisk (*) indicates the correct option when applicable.

Phenomena	**Illustrative Example**
argument-structure	Most cashiers are disliked*/flirted.
binding	Carlos said that Lori helped him*/himself.
determiner noun agreement	Craig explored that grocery store*/stores.
event structure	Give them to a library or burn them. ⇒ Distributive
filler-gap	Brett knew what many waiters find.*/Brett knew that many waiters find.
genericity	I assume you mean the crazy horse memorial. ⇒ Not Dynamic
island-effects	Which bikes is John fixing?*/Which is John fixing bikes?
antonym negation	It was not*/really hot, it was cold.
negative polarity item licensing	Only/Even Bill would ever complain.
semantic proto-roles	These look fine to me. ⇒ Exists as physical
quantifiers	There aren't many*/all lights darkening.
rhetorical structure theory	The statistics quoted by the " new " Census Bureau report ⇒ Elaboration
subject-verb agreement	A sketch of lights does not*/do not appear.

Table 3: List of linguistic phenomena and their corresponding subfields in the Holmes benchmark.

linguistic phenomena	subfield	linguistic phenomena	subfield
next sentence prediction	discourse	semantic odd man out	semantics
rhetorical structure theory	discourse	word sense	semantics
sentence order	discourse	word content	semantics
discourse connective	discourse	coordination inversion	semantics
coreference resolution	discourse	object animacy	semantics
bridging	discourse	event structure	semantics
irregular forms	morphology	factuality	semantics
subject-verb agreement	morphology	complex words	semantics
determiner noun agreement	morphology	genericity	semantics
anaphor agreement	morphology	metaphor	semantics
age comparison	reasoning	named entity labeling	semantics
negation	reasoning	negative polarity item licensing	semantics
speculation	reasoning	argument structure	syntax
multi-hop composition	reasoning	bigram-shift	syntax
property conjunction	reasoning	binding	syntax
object comparison	reasoning	tree-depth	syntax
antonym negation	reasoning	case	syntax
encyclopedic composition	reasoning	subject-verb agreement	syntax
taxonomy conjunction	reasoning	anaphor agreement	syntax
always never	reasoning	top-constituent-task	syntax
object gender	semantics	subject number	syntax
passive	semantics	deoncausative-inchoative alternation	syntax
protoroles	semantics	control / raising	syntax
quantifiers	semantics	ellipsis	syntax
synonym-/antonym-detection	semantics	sentence length	syntax
verb dynamic	semantics	filler gap	syntax
semantic role labeling	semantics	readability	syntax
sentiment analysis	semantics	island effects	syntax
time	semantics	local attractor	syntax
subject animacy	semantics	part-of-speech	syntax
subject gender	semantics	object number	syntax
tense	semantics	constituent parsing	syntax
relation classification	semantics	negative polarity item licensing	syntax

F BERT finetuning on Harry Potter dataset

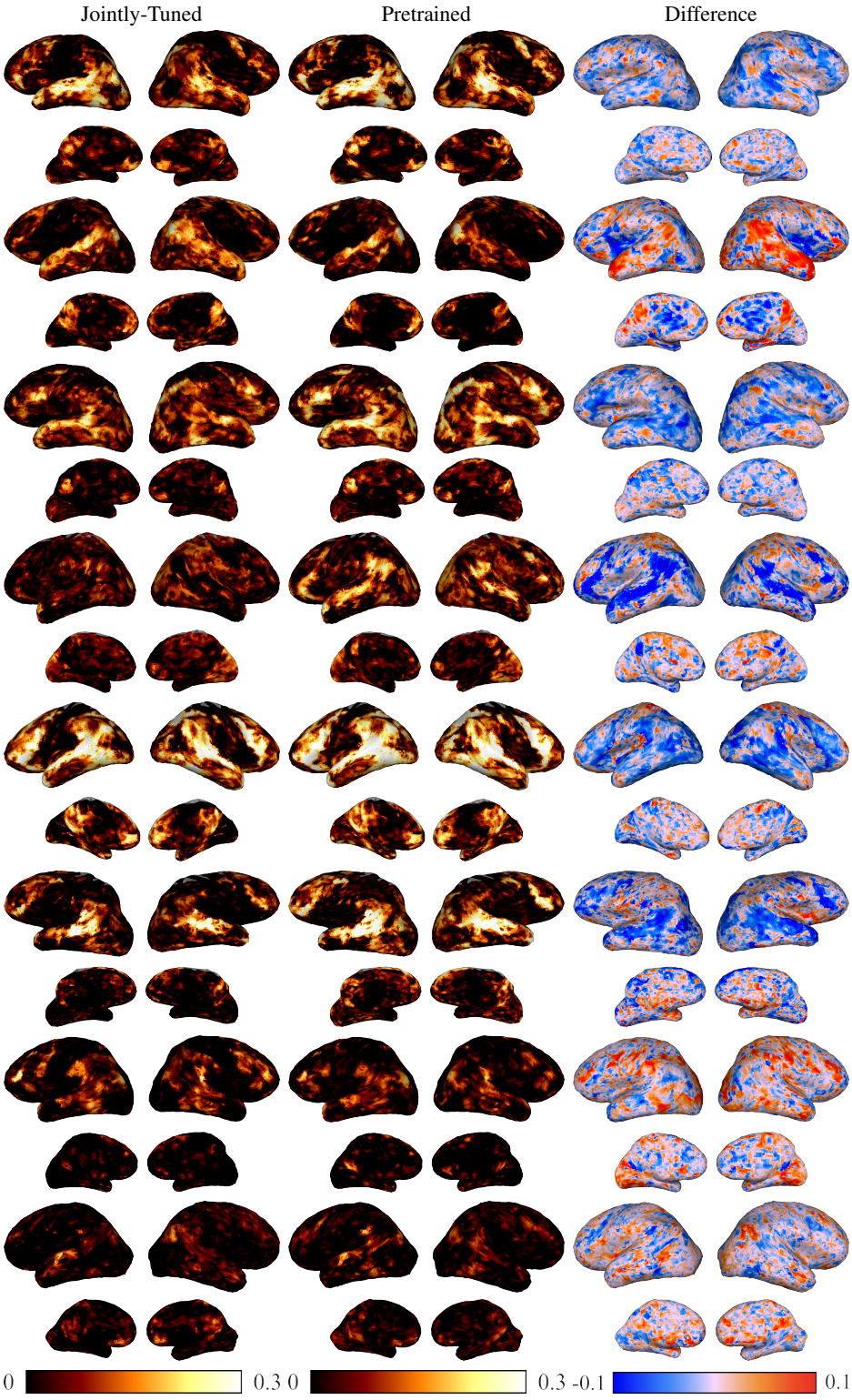

Figure 7: Performances of BERT-based Jointly-Tuned and pretrained models at the brain alignment task. Brain plots show voxel-wise Pearson correlations between model activations and brain responses for each subject on the Harry Potter dataset. The left column displays results for the Jointly-Tuned model, the center column for the pretrained model, and the right column shows their difference (Jointly-Tuned minus pretrained). Warmer colors indicate stronger alignment with brain activity. These results illustrate the distribution of brain alignment across subjects and highlight areas where brain jointly-tuning has effects.

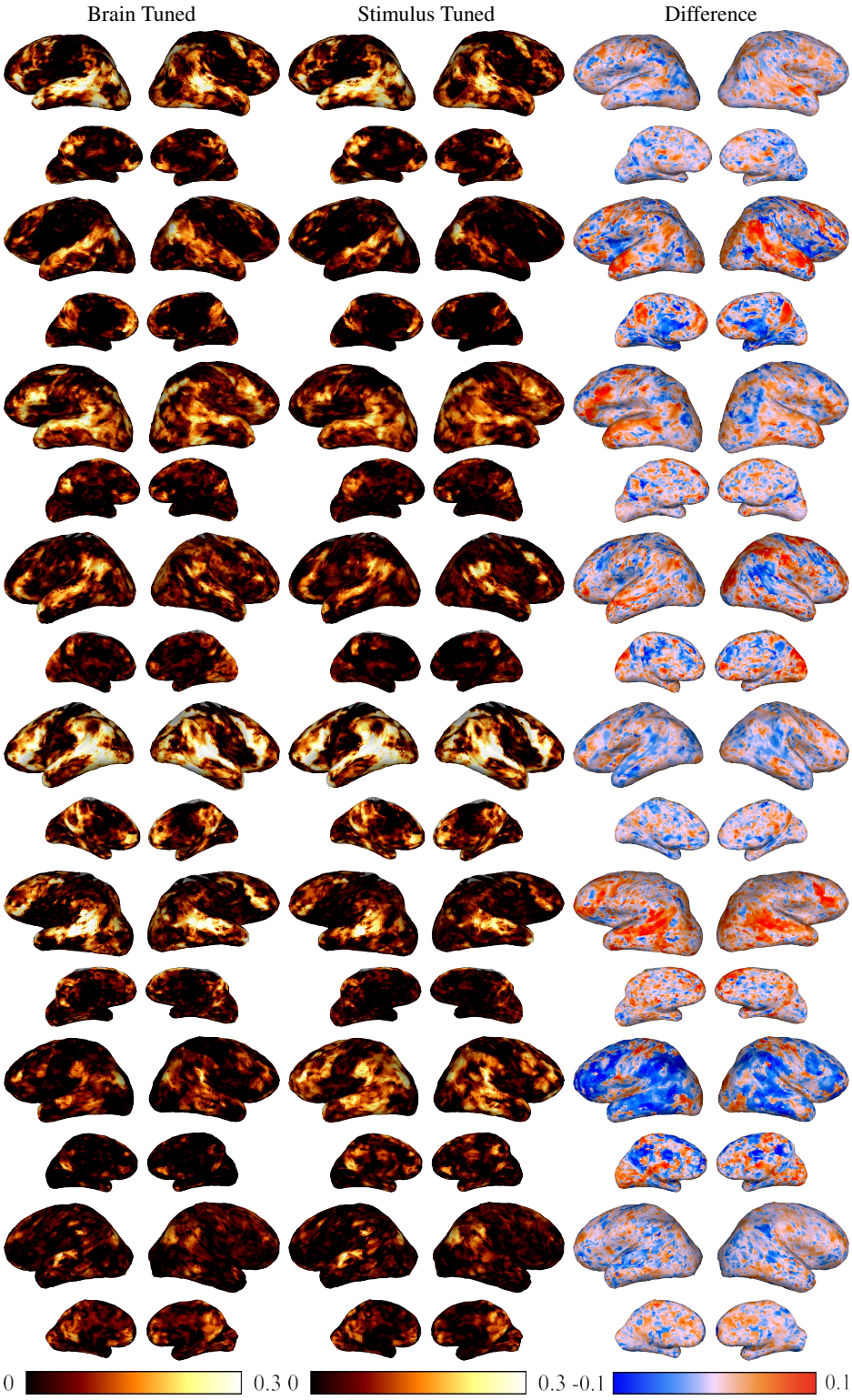

Figure 8: Performances of BERT-based Brain-Tuned and Stimulus-Tuned models on the Harry Potter dataset at the brain alignment task. Brain plots show voxel-wise Pearson correlations between model activations and brain responses for each subject. The left column displays results for the Brain-Tuned model, the center column for the Stimulus-Tuned model, and the right column shows their difference (Brain-Tuned minus Stimulus-Tuned). Warmer colors indicate stronger alignment with brain activity. These results illustrate the distribution of brain alignment across subjects and highlight areas where brain-tuning has effects.

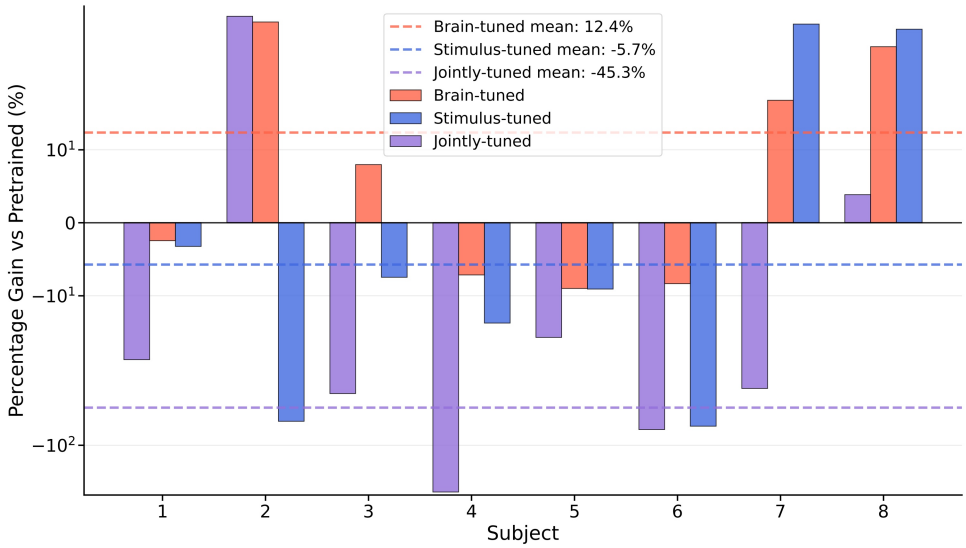

Figure 9: Average relative improvement and standard error of the brain alignment, measured using Pearson correlation in language-related ROIs filtered using noise ceiling values, for the BERT-based Jointly-Tuned, Brain-Tuned and Stimulus-Tuned models relative to the pretrained baseline. Each bar represents the percentage gain for an individual subject in the Harry Potter dataset (Wehbe et al., 2014a). Dashed lines indicate the mean gain across all participants. The Brain-Tuned models show a consistent average improvement in alignment compared to the pretrained BERT, while Stimulus-Tuned and Jointly-Tuned models perform worse on average.

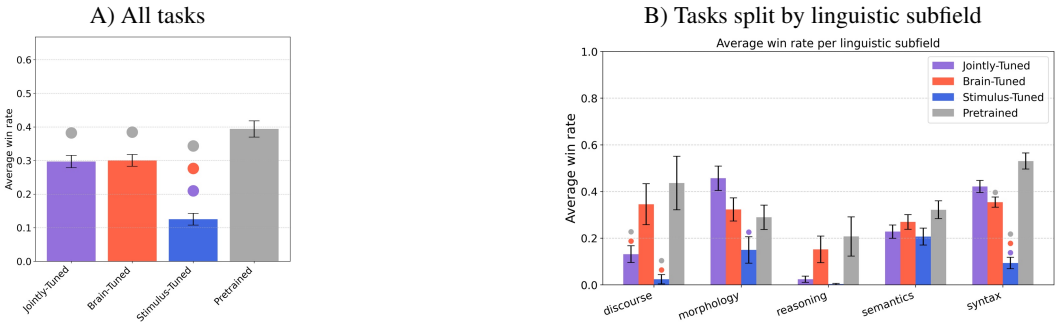

Figure 10: Average win rate and standard error of the BERT-based Brain-Tuned and Stimulus-Tuned models across participants and taks(Left) and across different linguistic subfields (Right). The win rate indicates how often each model outperforms its counterpart across tasks and participants. The Brain-Tuned models significantly outperforms the Stimulus-Tuned models ($p < 0.001$, indicated by ***), as assessed using a Wilcoxon signed-rank test (Left). This result suggests that enforcing brain alignment positively influences linguistic competence. The Brain-Tuneds model shows a higher win rate in all linguistic subfield (Right) and show significantly higher for syntax and reasoning subfields ($p < 0.05$, Wilcoxon signed-rank test with Holm-Bonferroni correction), suggesting that brain alignment particularly affect syntax and reasoning tasks.

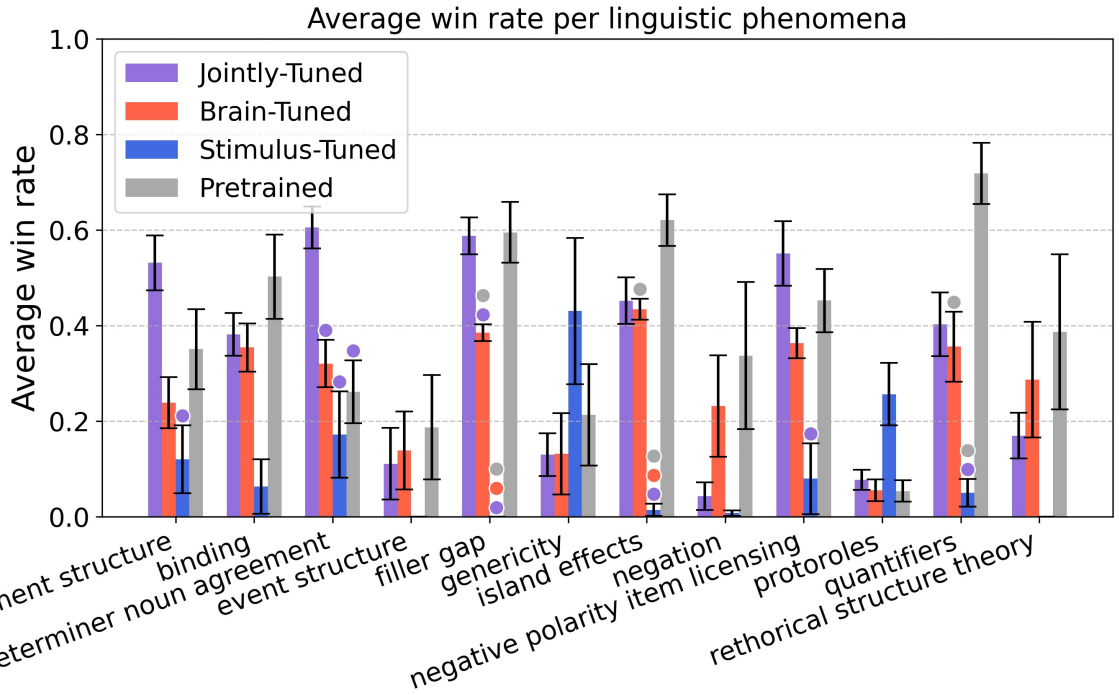

Figure 11: Average win rate with standard error across various linguistic phenomena for the GPT2- based Jointly-Tuned, Brain-Tuned, Stimulus-Tuned and pretrained models on the Harry Potter dataset. Each bar represents the average win rate for a specific linguistic phenomenon, with error bars indicating standard error. Some concrete examples of the linguistic tasks are provided in the Table2.

G GPT2 finetuning on Harry Potter dataset

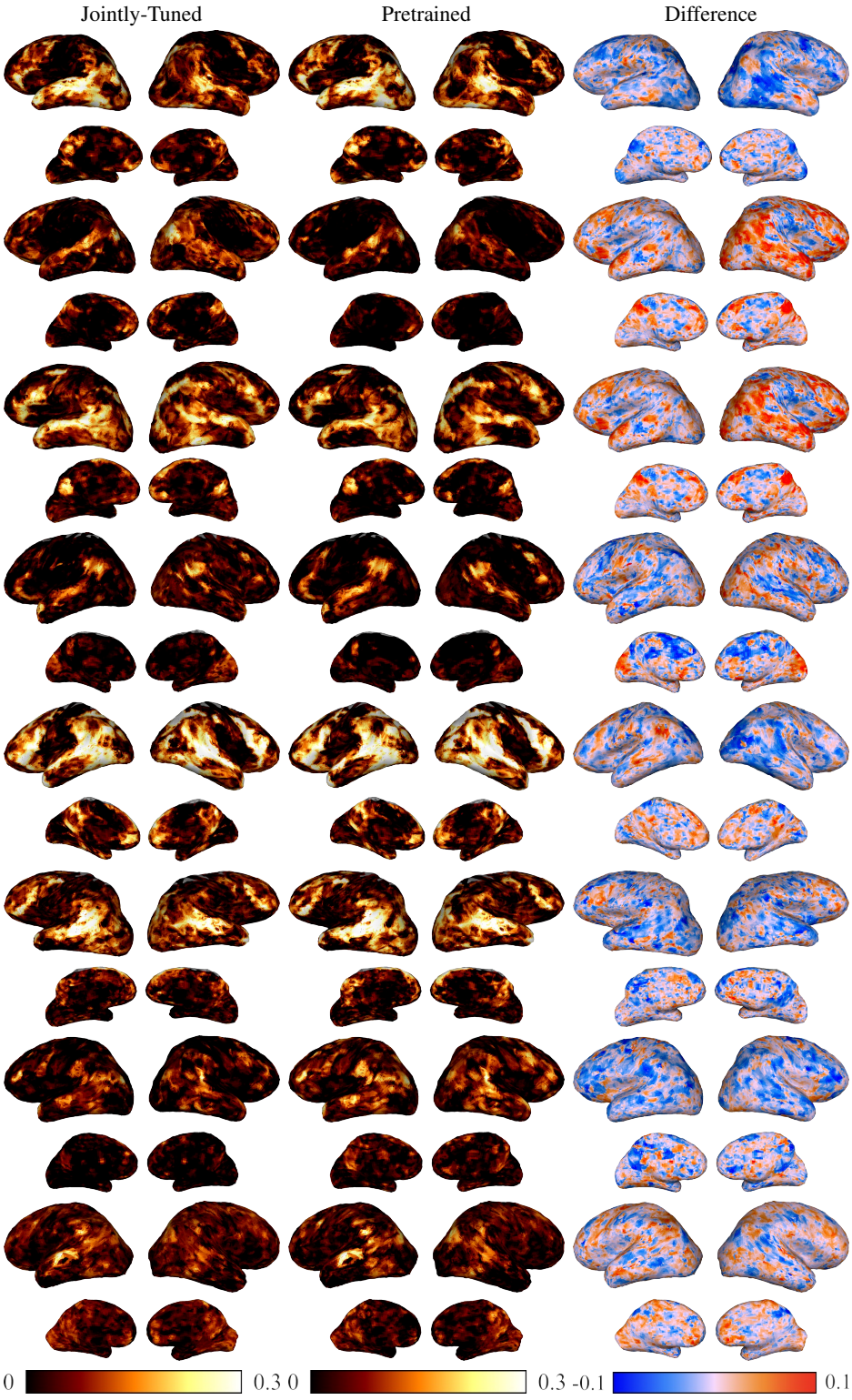

Figure 12: Performances of GPT2-based Jointly-Tuned and pretrained models at the brain alignment task. Brain plots show voxel-wise Pearson correlations between model activations and brain responses for each subject on the Harry Potter dataset. The left column displays results for the Jointly-Tuned model, the center column for the pretrained model, and the right column shows their difference (Jointly-Tuned minus pretrained). Warmer colors indicate stronger alignment with brain activity. These results illustrate the distribution of brain alignment across subjects and highlight areas where brain jointly-tuning has effects.

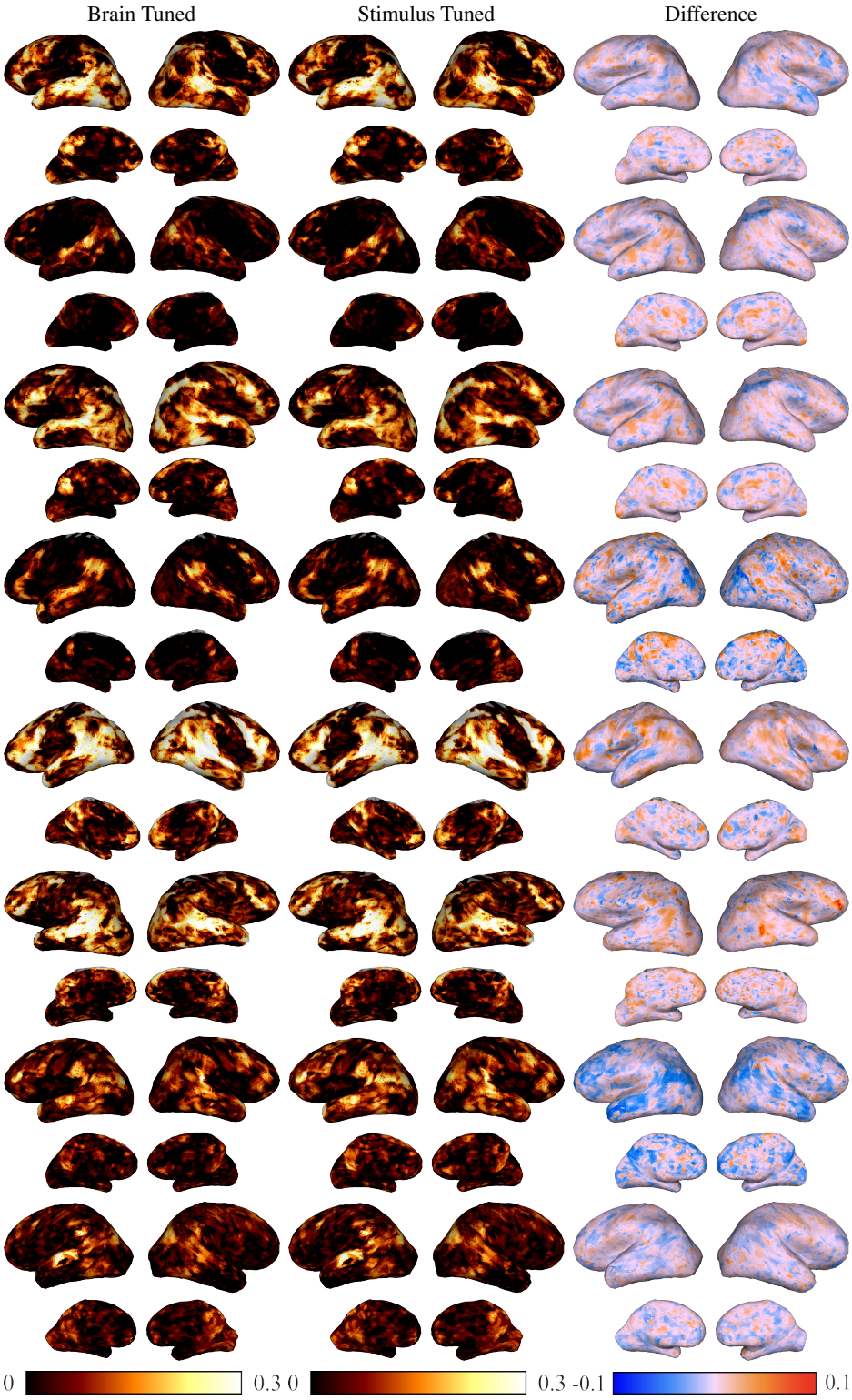

Figure 13: Performances of GPT2-based Brain-Tuned and Stimulus-Tuned models on the Harry Potter dataset at the brain alignment task. Brain plots show voxel-wise Pearson correlations between model activations and brain responses for each subject. The left column displays results for the Brain-Tuned model, the center column for the Stimulus-Tuned model, and the right column shows their difference (Brain-Tuned minus Stimulus-Tuned). Warmer colors indicate stronger alignment with brain activity. These results illustrate the distribution of brain alignment across subjects and highlight areas where brain-tuning has effects.

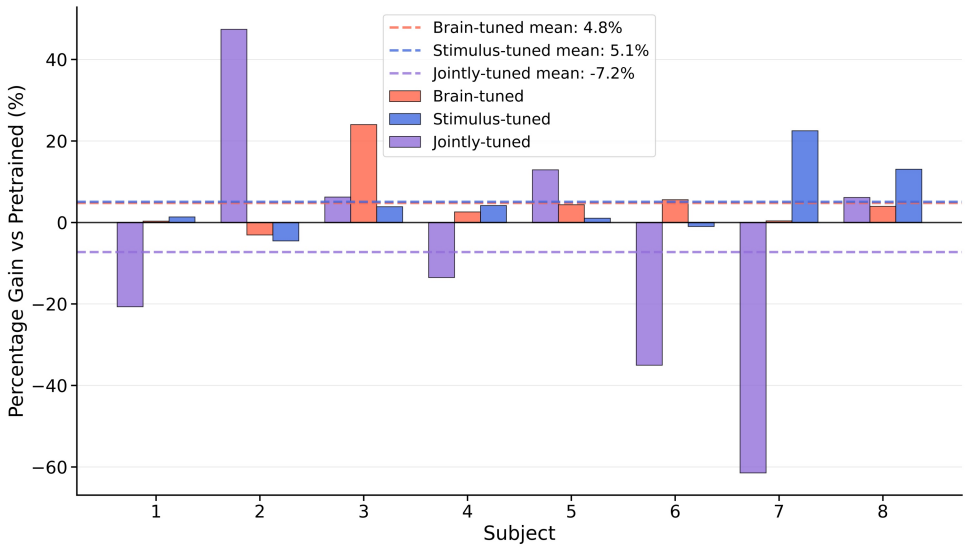

Figure 14: Average relative improvement and standard error of the brain alignment, measured using Pearson correlation in language-related ROIs filtered using noise ceiling values, for the GPT2-based Jointly-Tuned, Brain-Tuned and Stimulus-Tuned models relative to the pretrained baseline. Each bar represents the percentage gain for an individual subject in the Harry Potter dataset (Wehbe et al., 2014a). Dashed lines indicate the mean gain across all participants. The Brain-Tuned models show an average improvement similar to the Stimulus-Tuned models, while the Jointly-Tuned models show worse alignment on average.

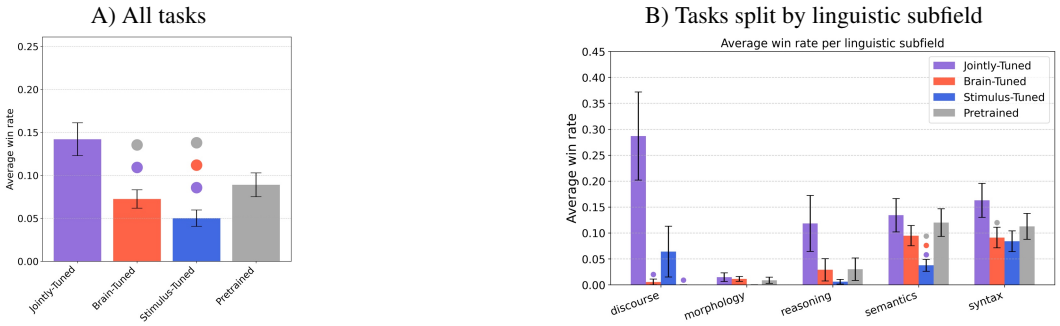

Figure 15: Average win rate and standard error of the GPT2-based Brain-Tuned and Stimulus-Tuned models across participants and taks(Left) and across different linguistic subfields (Right). The win rate indicates how often each model outperforms its counterpart across tasks and participants. The Brain-Tuned models significantly outperforms the Stimulus-Tuned models ($p < 0.001$, indicated by *), as assessed using a Wilcoxon signed-rank test (Left). This result suggests that enforcing brain alignment positively influences linguistic competence. The Brain-Tuned models shows a higher win rate in all linguistic subfield (Right) and show significantly higher for syntax and reasoning subfields ($p < 0.05$, Wilcoxon signed-rank test with Holm-Bonferroni correction), suggesting that brain alignment particularly affect syntax and reasoning tasks.

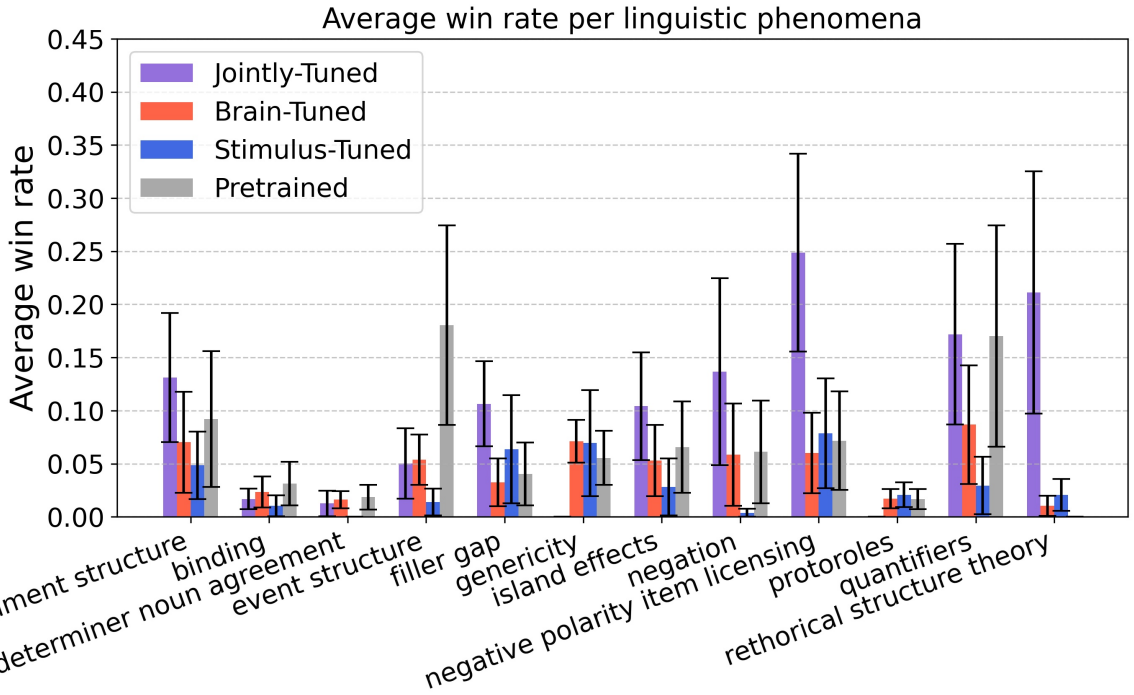

Figure 16: Average win rate with standard error across various linguistic phenomena for the GPT2- based Jointly-Tuned, Brain-Tuned, Stimulus-Tuned and pretrained models on the Harry Potter dataset. Each bar represents the average win rate for a specific linguistic phenomenon, with error bars indicating standard error. Some concrete examples of the linguistic tasks are provided in the Table2.

H BERT finetuning on Moth Radio Hour dataset

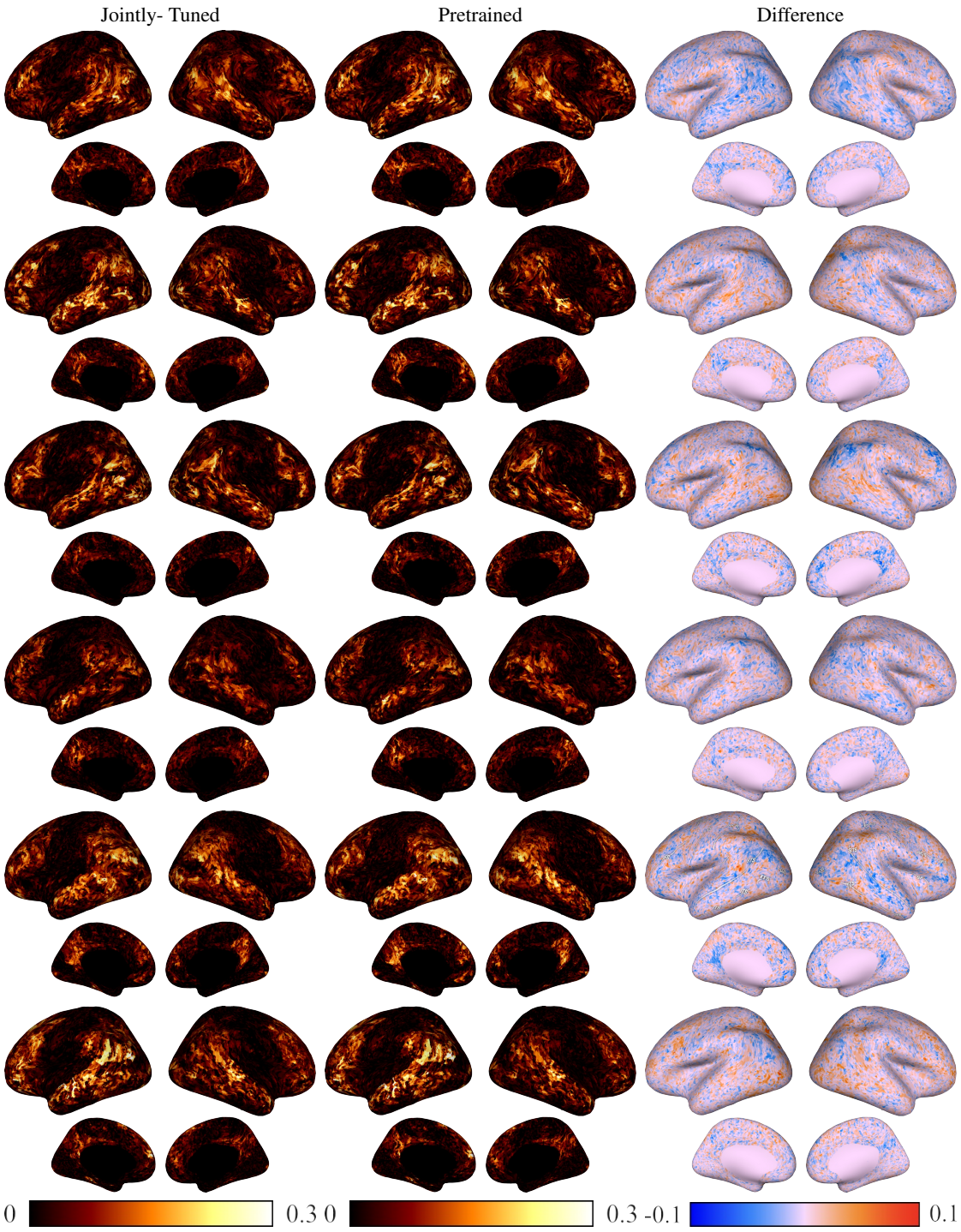

Figure 17: Performances of BERT-based Jointly-Tuned and pretrained models at the brain alignment task. Brain plots show voxel-wise Pearson correlations between model activations and brain responses for each subject on the Moth Radio Hour dataset. The left column displays results for the Jointly-Tuned model, the center column for the pretrained model, and the right column shows their difference (Jointly-Tuned minus pretrained). Warmer colors indicate stronger alignment with brain activity. These results illustrate the distribution of brain alignment across subjects and highlight areas where brain jointly-tuning has effects.

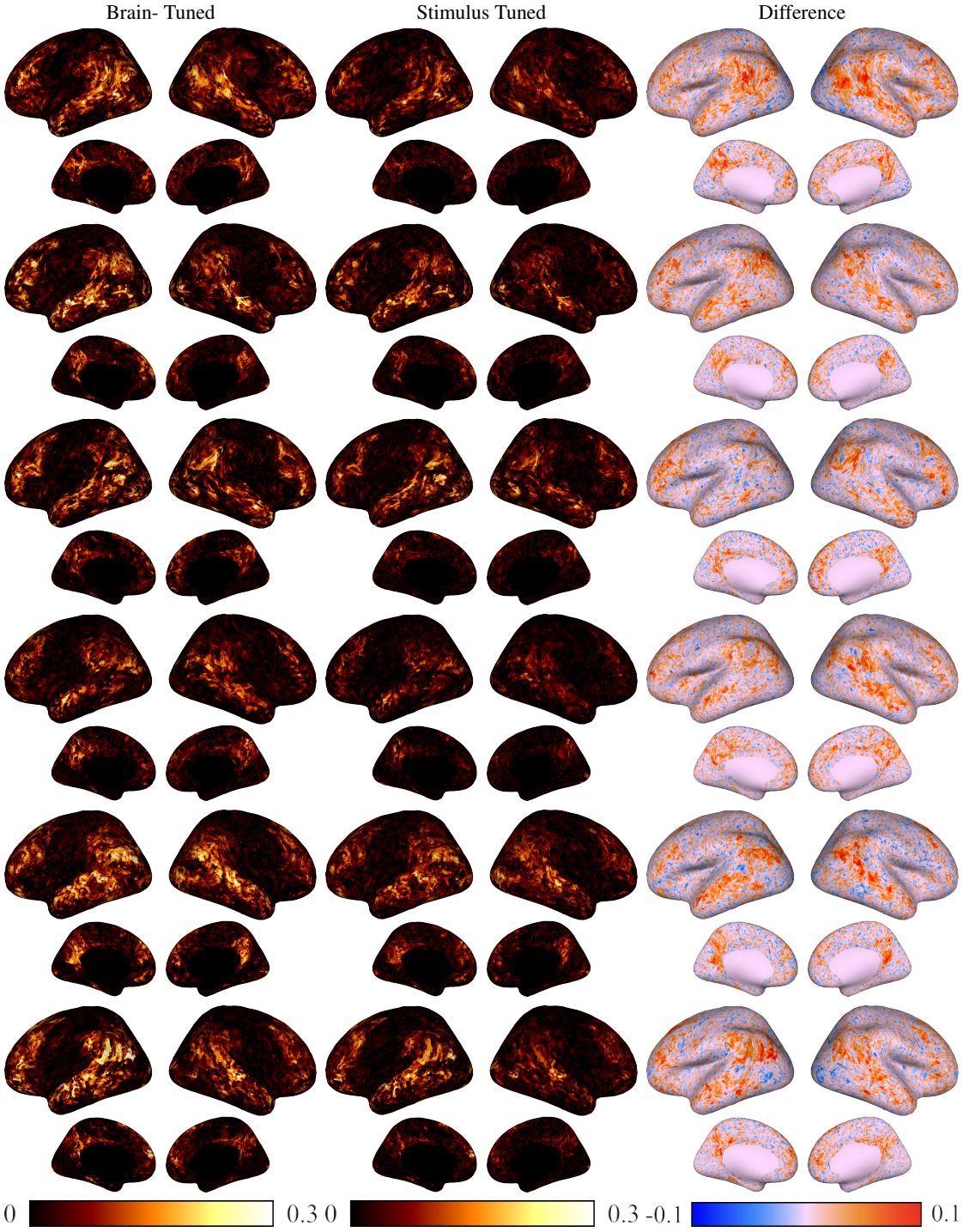

Figure 18: Performances of BERT-based Brain-Tuned and Stimulus-Tuned models on the Moth Radio Hour dataset at the brain alignment task. Brain plots show voxel-wise Pearson correlations between model activations and brain responses for each subject. The left column displays results for the Brain-Tuned model, the center column for the Stimulus-Tuned model, and the right column shows their difference (Brain-Tuned minus Stimulus-Tuned). Warmer colors indicate stronger alignment with brain activity. These results illustrate the distribution of brain alignment across subjects and highlight areas where brain-tuning has effects.

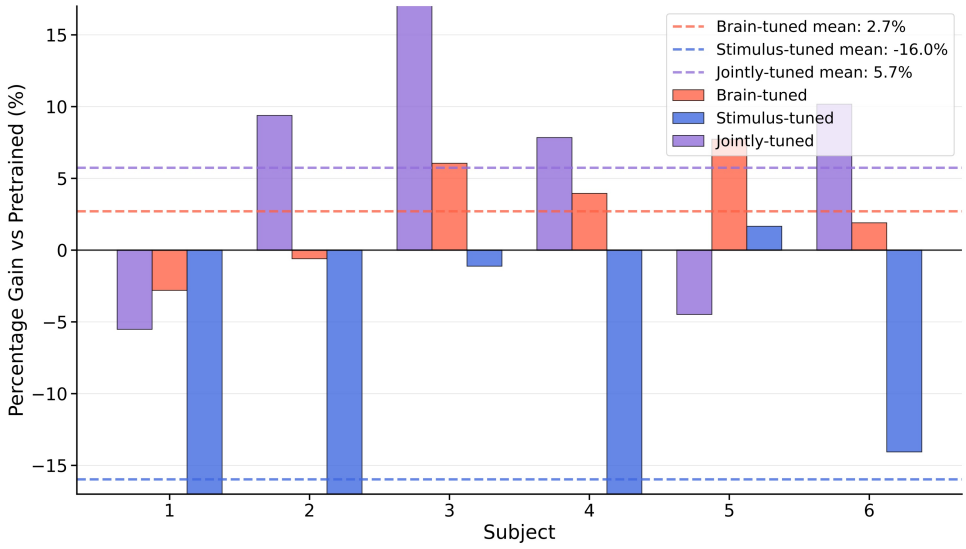

Figure 19: Average relative improvement and standard error of the brain alignment, measured using Pearson correlation in language-related ROIs filtered using noise ceiling values, for the BERT-based Jointly-Tuned, Brain-Tuned and Stimulus-Tuned models relative to the pretrained baseline. Each bar represents the percentage gain for an individual subject in The Moth Radio Hour dataset (Deniz et al., 2019). Dashed lines indicate the mean gain across all participants. Both Brain-Tuned and Jointly-Tuned models show a consistent average improvement in alignment compared to the pretrained BERT, while the Stimulus-Tuned models perform worse on average.

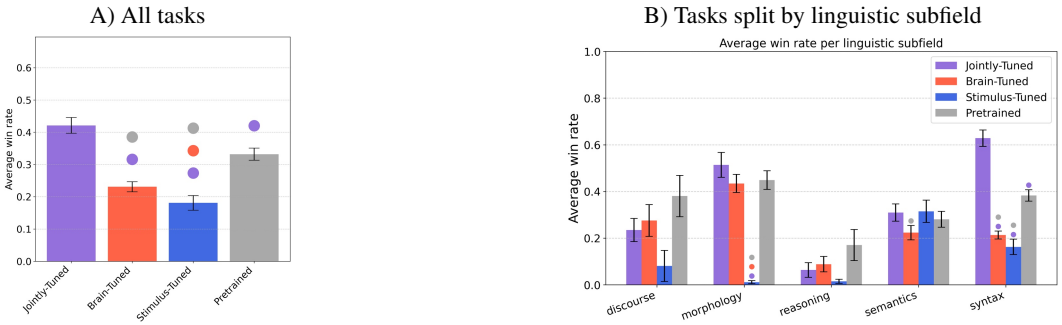

Figure 20: Average win rate and standard error of the BERT-based Jointly-Tuned, Brain-Tuned, Stimulus-Tuned, and pretrained models across participants and tasks (Left) and across different linguistic subfields (Right) on The Moth Radio Hour dataset. The win rate indicates how often each model outperforms its counterpart across tasks and participants. For each bar, the coloured dots on top indicate models that perform significantly better (assessed using a Wilcoxon signed-rank test). The Jointly-Tuned model is significantly better than the pretrained across all tasks, and the Brain-Tuned model is significantly better than the Stimulus-Tuned model (Left). This indicates that joint fine-tuning improves linguistic competence, and training on brain data is more effective than fine-tuning with stimuli. The Jointly-Tuned model shows a higher win rate for syntax, semantics and morphology linguistic subfield, and significantly higher for syntax (as assessed with a Wilcoxon signed-rank test with Holm-Bonferroni correction), while the Brain-Tuned model performs better than the Stimulus-Tuned model for almost all linguistic subfields and significantly higher for morphology(Right), suggesting that brain alignment particularly affects morphology tasks.

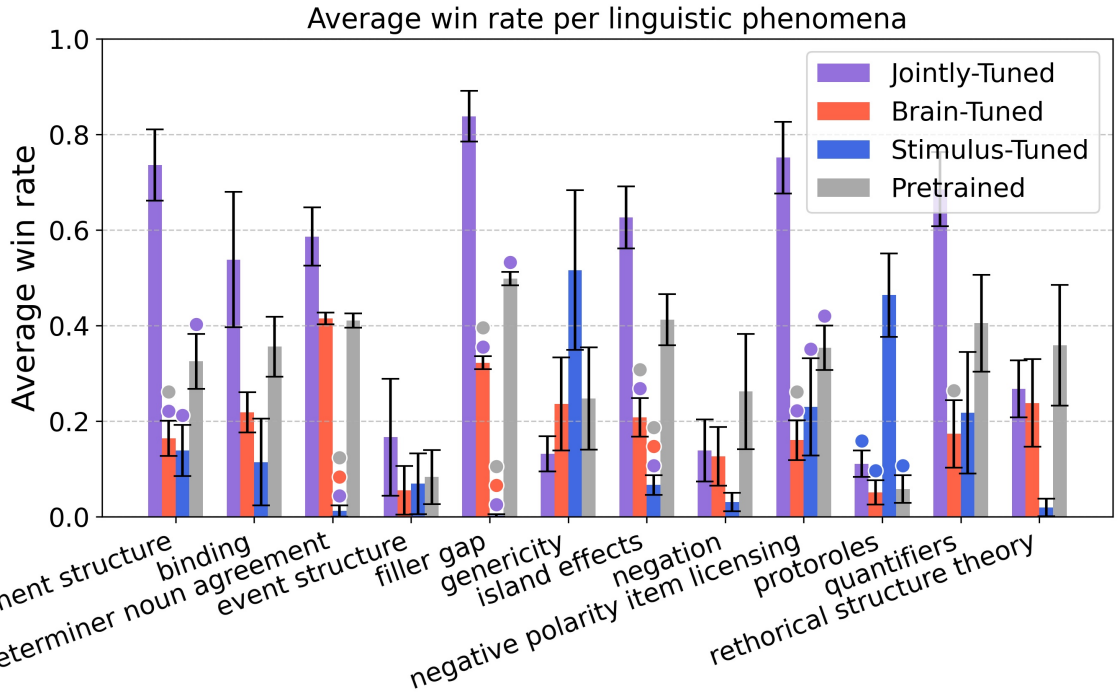

Figure 21: Average win rate with standard error across various linguistic phenomena for the BERT- based Jointly-Tuned, Brain-Tuned, Stimulus-Tuned and pretrained model on the Moth Radio Hour dataset. Each bar represents the average win rate for a specific linguistic phenomenon, with error bars indicating standard error. Some concrete examples of the linguistic tasks are provided in the Table2.

I GPT2 finetuning on Moth Radio Hour dataset

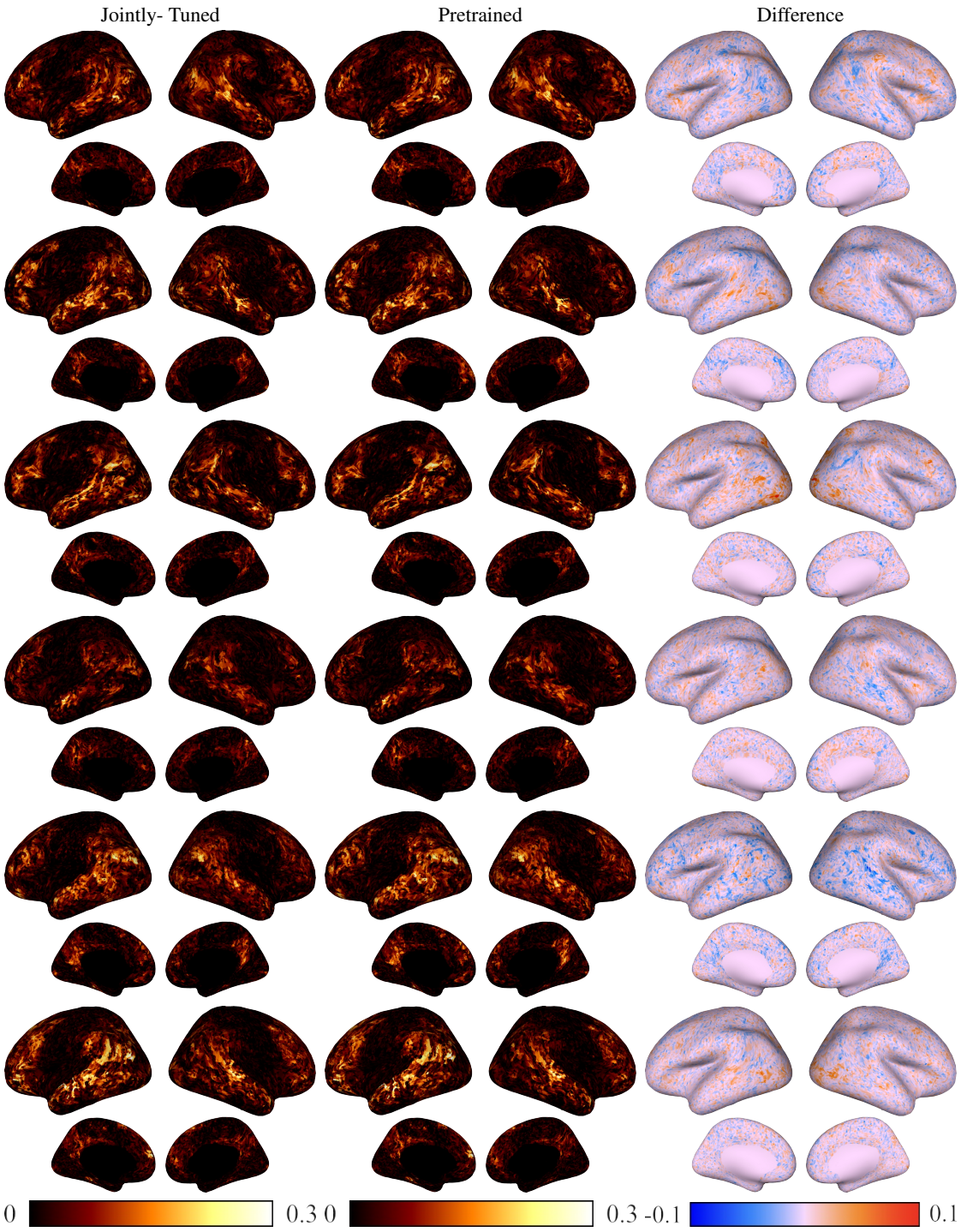

Figure 22: Performances of GPT2-based Jointly-Tuned and pretrained models at the brain alignment task. Brain plots show voxel-wise Pearson correlations between model activations and brain responses for each subject on the Moth Radio Hour dataset. The left column displays results for the Jointly-Tuned model, the center column for the pretrained model, and the right column shows their difference (Jointly-Tuned minus pretrained). Warmer colors indicate stronger alignment with brain activity. These results illustrate the distribution of brain alignment across subjects and highlight areas where brain jointly-tuning has effects.

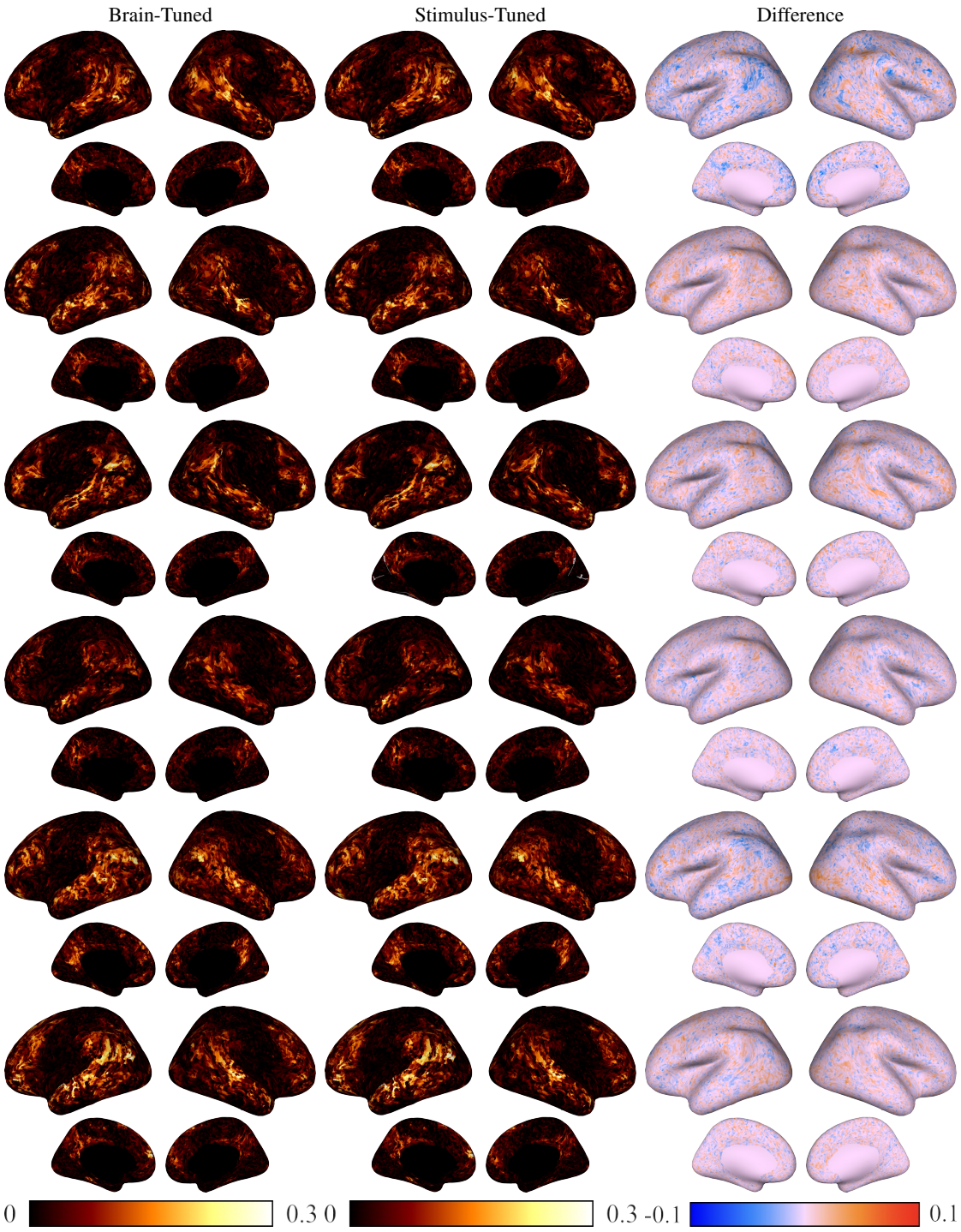

Figure 23: Performances of GPT2-based Brain-Tuned and Stimulus-Tuned models on the Moth Radio Hour dataset at the brain alignment task. Brain plots show voxel-wise Pearson correlations between model activations and brain responses for each subject. The left column displays results for the Brain-Tuned model, the center column for the Stimulus-Tuned model, and the right column shows their difference (Brain-Tuned minus Stimulus-Tuned). Warmer colors indicate stronger alignment with brain activity. These results illustrate the distribution of brain alignment across subjects and highlight areas where brain-tuning has effects.

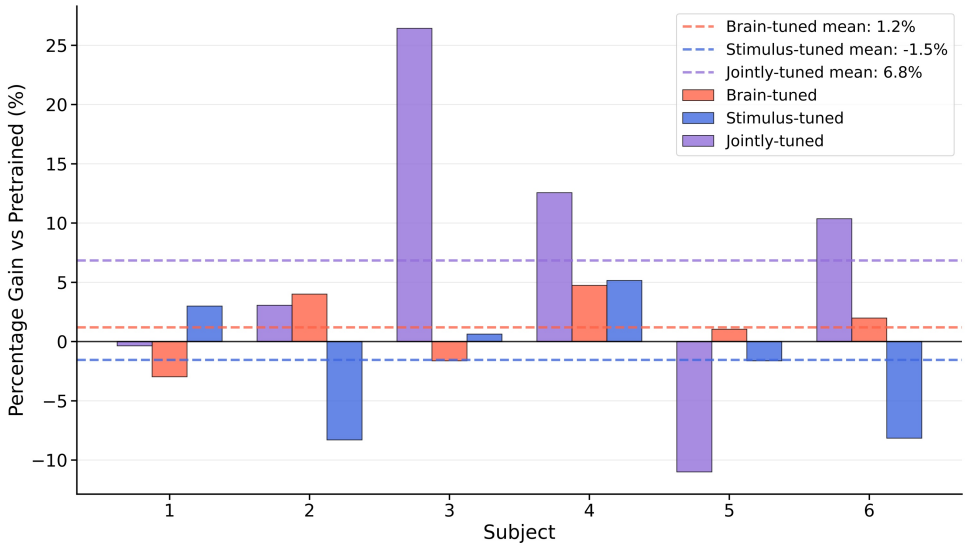

Figure 24: Average relative improvement and standard error of the brain alignment, measured using Pearson correlation in language-related ROIs filtered using noise ceiling values, for the GPT2-based Jointly-Tuned, Brain-Tuned and Stimulus-Tuned models relative to the pretrained baseline. Each bar represents the percentage gain for an individual subject in The Moth Radio Hour dataset (Deniz et al., 2019). Dashed lines indicate the mean gain across all participants. Both Brain-Tuned and Jointly-Tuned models show a consistent average improvement in alignment compared to the pretrained GPT2, while the Stimulus-Tuned models perform worse on average.

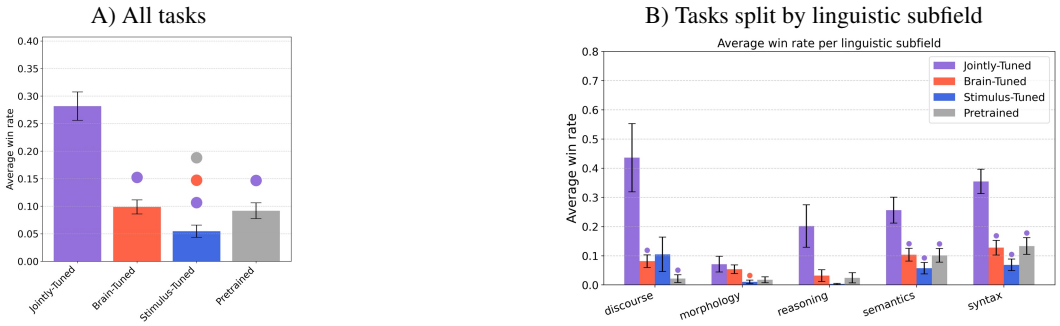

Figure 25: Average win rate and standard error of the GPT2-based Jointly-Tuned, Brain-Tuned, Stimulus-Tuned, and pretrained models across participants and tasks (Left) and across different linguistic subfields (Right) on The Moth Radio Hour dataset. The win rate indicates how often each model outperforms its counterpart across tasks and participants. For each bar, the coloured dots on top indicate models that perform significantly better (assessed using a Wilcoxon signed-rank test). The Jointly-Tuned model is significantly better than the pretrained across all tasks, and the Brain-Tuned model is significantly better than the Stimulus-Tuned model (Left). This indicates that joint fine-tuning improves linguistic competence, and training on brain data is more effective than fine-tuning with stimuli. The Jointly-Tuned model shows a higher win rate for every linguistic subfield, and significantly higher for discourse, semantics and syntax (as assessed with a Wilcoxon signed-rank test with Holm-Bonferroni correction), while the Brain-Tuned model performs better than the Stimulus-Tuned model for almost all linguistic subfields and significantly higher for morphology(Right), suggesting that brain alignment particularly affects morphology tasks.

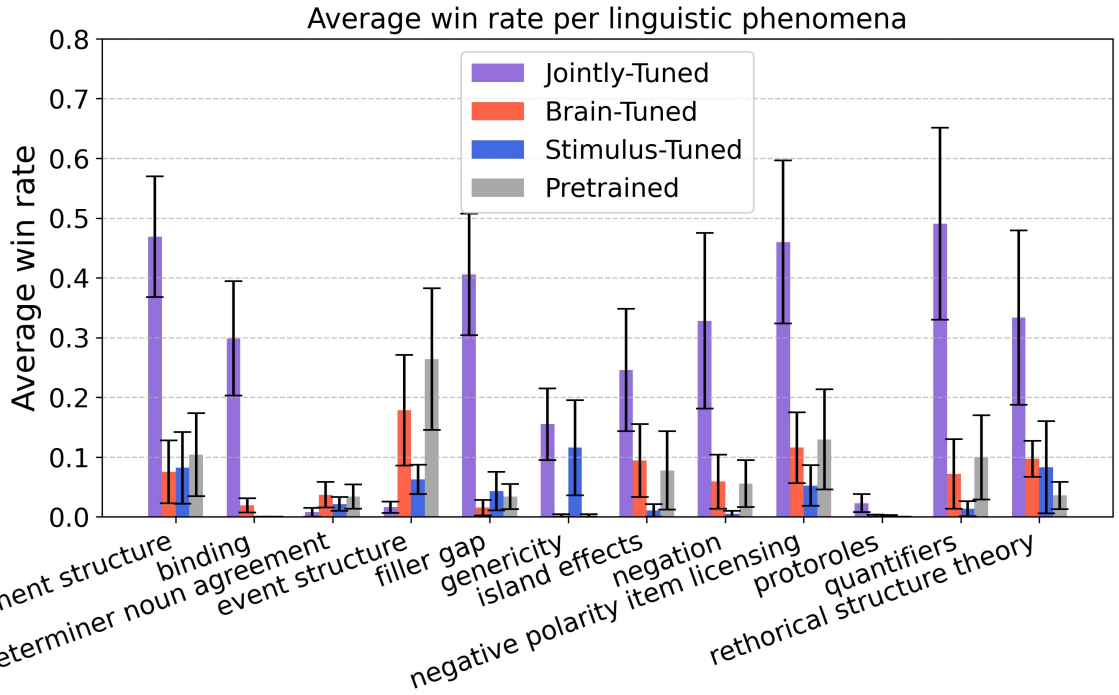

Figure 26: Average win rate with standard error across various linguistic phenomena for the GPT2- based Jointly-Tuned, Brain-Tuned, Stimulus-Tuned and pretrained model on the Moth Radio Hour dataset. Each bar represents the average win rate for a specific linguistic phenomenon, with error bars indicating standard error. Some concrete examples of the linguistic tasks are provided in the Table2.