We thank the reviewers for the thoughtful suggestions and attempt to address their questions within the space constraints.

All reviewers: To better interpret our results, we have a new analysis using additional labels released with the Harry Potter data which identify the presence of various syntactic, semantic, and emotional features for each word in the chapter. We score each input example of 20 words as to how much fine-tuning hurts or harms the example. On manually selected language-region voxels, we compute the difference in the distance from the model prediction to the target between the fine-tuned and vanilla BERT models (for our best participant). We compare the distributions of the features on the examples most helped and most harmed by fine-tuning, as determined by this metric, and find some indications that features related to emotion, the subject dependency-role, and noun representations are improved by fine-tuning. We will present this analysis in the main paper, and we think that better understanding the changes in the model will be an exciting area for future research.

R2 and R3: Our use of the 20 vs. 20 evaluation follows previous work using this dataset (cf. Wehbe *et al.* 2014a, 2014b in the paper). In this experiment the text is shown to the participant only once, so the SNR is very low. 20 vs. 20 boosts the SNR without the averaging that is normally used in a multiple repetition setting. It enables us to compare models more easily than R^2 which is dominated by noise. Qualitatively the brain maps of prediction performance in our model comparisons look similar using either metric, and we will add the R^2 maps as a supplementary figure.

We agree that we needed to quantify the results in fig. 2. For the models where it was computationally feasible (all but the fully jointly trained model) we trained the models 100 times ($25 \times \langle 4 \text{ CV folds} \rangle$) with different initializations. The models all use the same initialization for run i so we use a paired t-test per voxel to evaluate whether voxel prediction accuracies are different between two models, correcting for false discovery rate at a .01 level (Benjamini 1995 JRoyalStatSoc). We plan to replace fig. 2 in the main paper with fig. 4 (a-c) for the models where we can, and to replace fig. 3 in the main paper with statistical maps similar to fig. 4 (f-i) here. Fig. 4 (f-i) shows that while the MEG to fMRI transfer learning does not appear to improve voxel prediction on average (fig. 4 (c), for almost every participant it helps prediction in language regions and harms prediction elsewhere (compare (f-i) with (d)). The harm is likely due to overfitting.

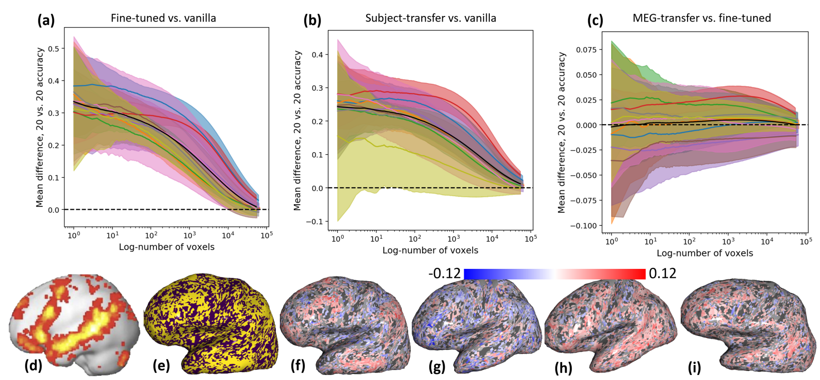

Figure 4: **(a-c):** Each sub-figure compares two models. Voxels are sorted on the x-axis in descending order of the maximum of the two models' accuracies. The colored lines (one per participant) show mean accuracies taken over all voxels to the left of each x-coord. Shaded regions show std. deviation over 100 model initializations. The black line is the mean over all participants. **(d):** Regions typically associated with language processing (adapted from Fedorenko *et al.* 2012 Neuropsychologia). **(e):** An example significance mask. **(f-i):** Comparison of the MEG-transfer and fine-tuned models (left hemispheres). Voxels are masked out where differences are not significant and colored according to the mean difference in accuracy between the MEG transfer and fine-tuned models. Notice the bottom parts of plots (e-f) look different than (d) because they don't include the cerebellum.

This outcome also relates to motivation. It is not self-evident that fine-tuning should help prediction. It is possible that there could be nothing to learn beyond what is encoded in BERT (i.e. for vanilla to be the best possible fit even in a setting with a large number of samples), or for overfitting to be too problematic in a practical setting (with a small number of samples). We demonstrate that prediction of language areas improves while prediction elsewhere does not.

R1 and R4: Thank you for the positive evaluation. We will add more detail about the CLS token as per R4's suggestion.

R2: When we run GLUE, we use normal inputs (i.e. we do not use 20-word inputs). A bidirectional model is only problematic if we are attempting to model how information transformations happen in the brain algorithmically. Here we are interested in nudging the information content in BERT to be similar to the brain, but not in making its algorithm similar to the brain. Since humans have more real word knowledge from which to make predictions as they read left-to-right — helping in tasks like anaphora resolution — it's plausible for a bidirectional model to be more similar to human representations by using right-to-left processing to make up for its lack of knowledge.

R3: We agree this paper was light on framing, we have attempted to provide some backing in the short space here but we will appropriately motivate the problem in the main paper. Please see above for suggested interpretation. Clarification: we do not claim to be improving on the vanilla BERT w.r.t. GLUE, only that we are not impairing performance.