A Supplemental figures

Figure 8 shows the change in brain decoding performance after fine-tuning grouped by subject, under both the mean squared error and rank metrics.

Figure 9 shows the change in brain decoding performance after fine-tuning grouped by model, under both the mean squared error and rank metrics.

B Hyperparameters

Hyperparameter	Value
Batch size	32
Learning rate	2×10^{-5}
LR warmup proportion	10%
Maximum sequence length	128
Fine-tuning steps	250

Table 3: Fine-tuning hyperparameters, shared across fine-tuning runs for all tasks. These hyperparameters mostly follow the suggestions of Devlin et al. (2018).

Hyperparameter	Value
Training epochs	10
Loss metric	L1
Maximum rank of B	30
Positive semi-definite?	Yes

Table 4: Syntactic probe hyperparameters (Section 3.2.2), following the defaults of Hewitt and Manning (2019).

C Custom task information

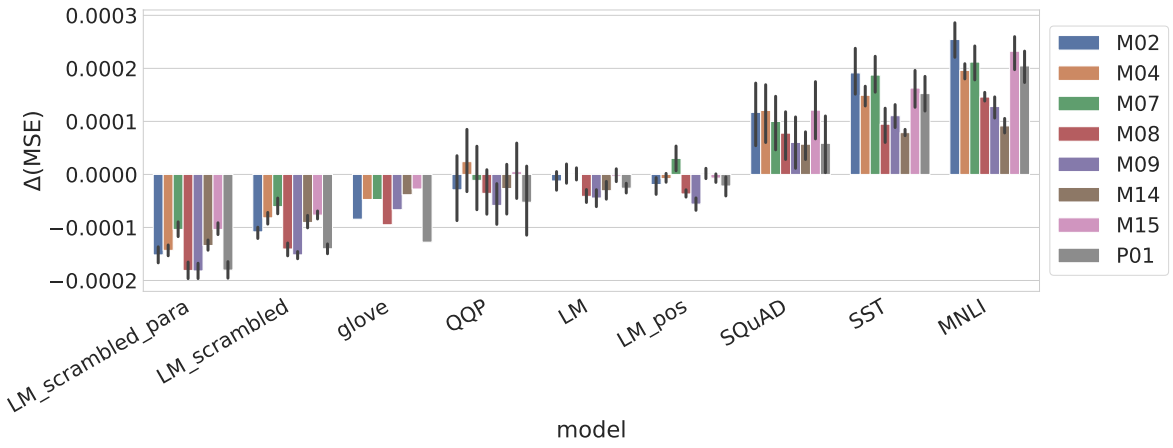
For each language modeling task, we randomly sampled and concatenated documents from the Toronto Books Corpus (Kiros et al., 2015). Each language modeling dataset contained 1,000,000 training sentences and 100,000 development and test sentences. (We generated over-sized datasets in order to ensure that multiple runs of the same model would be highly unlikely to see similar samples of training data.)

For the part-of-speech task, we tagged each sentence using spaCy (Honnibal and Montani, 2017) and followed the same random masking procedure as in the typical cloze language modeling task. spaCy assigned 49 unique part-of-speech tags to the sentences, yielding a 49-way classification task.

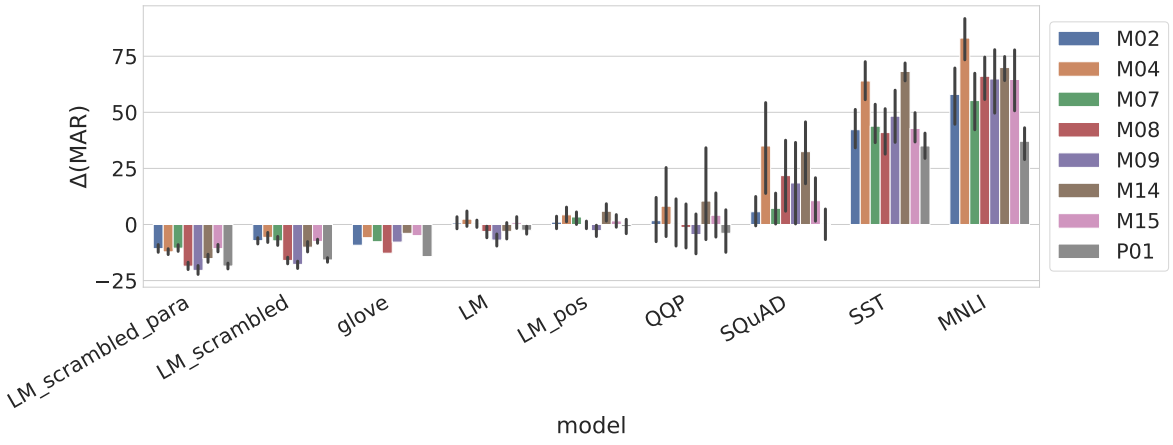
For all tasks, we retained the secondary BERT objective requiring the model to predict whether two sentences are adjacent or not in a source document. (This objective did not differ from the standard setup for the part-of-speech task; for the scrambling task, the input sentences were independently randomly shuffled.)

Table 5 shows training examples from each of these custom tasks. Figure 10 shows learning curves and validation accuracy curves for the models trained on each task.

Figure 8: Within-subject changes in brain decoding performance for different models, relative to the subject's brain decoding performance with the pre-trained BERT model. Error bars are 95% CIs pooling across decoders learned for different runs of each model (up to 8 per model).

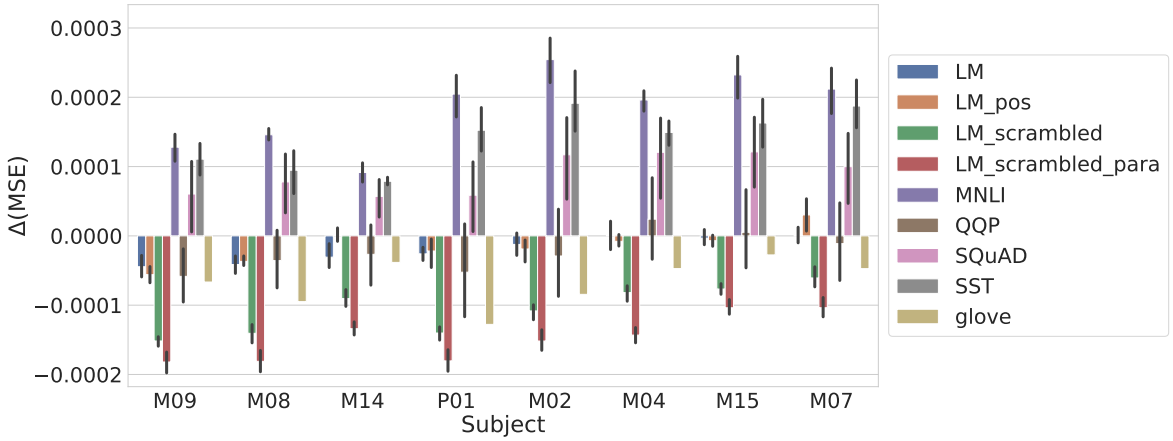


(a) Mean squared error metric

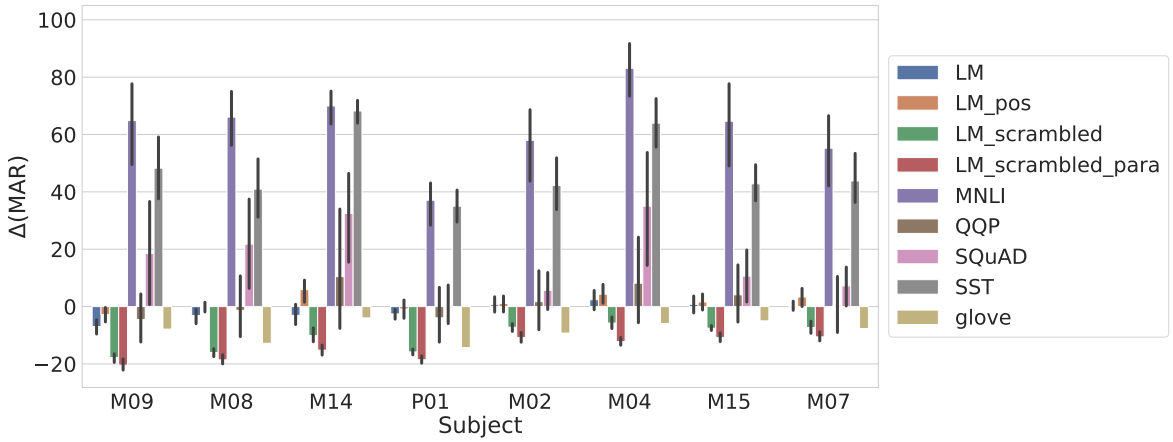


(b) Average rank metric

Figure 9: Within-model changes in brain decoding performance for different subjects, relative to the corresponding subject's brain decoding performance with the pre-trained BERT model. Error bars are 95% CIs pooling across decoders learned for different runs of each model (up to 8 per model).

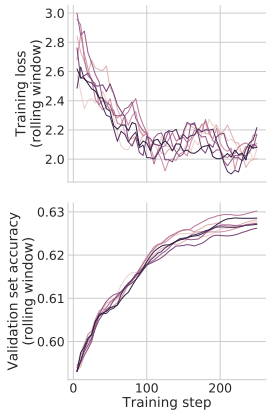


(a) Mean squared error metric

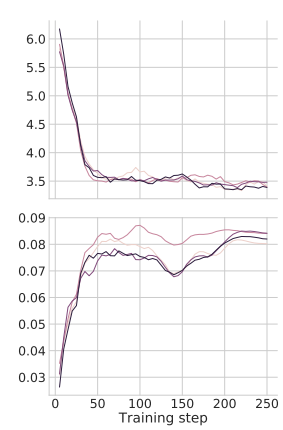


(b) Average rank metric

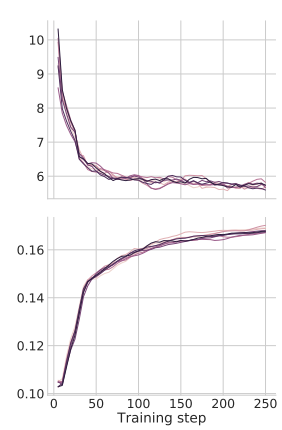
Figure 10: Learning curves (training set loss and evaluation set accuracy) for custom language modeling tasks.



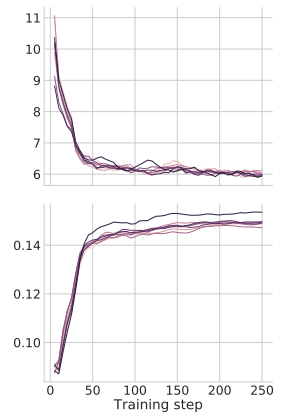
(a) LM (standard BERT task). Chance accuracy: $\frac{1}{30522} \approx 3 \times 10^{-5}$.



(b) LM-pos. Chance accuracy: $\frac{1}{49} \approx 0.02$.



(c) LM-scrambled. Chance accuracy: $\frac{1}{30522} \approx 3 \times 10^{-5}$.



(d) LM-scrambled-para. Chance accuracy: $\frac{1}{30522} \approx 3 \times 10^{-5}$.

Table 5: Examples from custom language modeling tasks.

Input	Ground-truth output
[MASK] minutes in she began to cry .	two
the door opened and lilith [MASK] in with worry on her face .	walked
she grabbed several pairs of jeans and some t - shirts [MASK] her closet	from

(a) LM

Input	Ground-truth output
, and [MASK] so drive can .	i
instead ##k kn documents pots important dish water be ##s greasy , [MASK] were of of and soon legal in dirty ##ick personal ##ks high to a bath and looking pan awaiting the i with ago pan lifestyle face many ##ils d , years .	piles
by connected will if you the ##tell we the that ##i [MASK] us lead ?	people
, and i so drive can . abraham want , i myself i [MASK] cars and whenever his , two , dad	have

(b) LM-scrambled

Input	Ground-truth output
scent sweat ta very [MASK] aroma ##ed caught , , a close a his . ##ana cabaret sand male skin she ##wny of up and ##al mu ##ting slick faint ...	##wood
, no in ! themselves supper with eyes included his ‘ , [MASK] began . the the alone everyone cried enjoying standing ...	triumph
, she told off out would wash followed “ asked her i with ” take . voice [MASK] ...	stepped

(c) LM-scrambled-para

Input	Ground-truth output
you ’ ve probably just gotten yourself off schedule a few days [MASK] you’ve been so busy and stressed .	`CD`
a couple of lordlings [MASK] .	`VBD`
the blond boy , who at fourteen was [MASK] much taller than anderra ... showed no signs of discomposure .	`RB`

(d) LM-pos