General response to reviewers: We thank the reviewers for their insightful and useful feedback! We are encouraged that they believe our work "opens up many possibilities" (**R3**), and is "very novel" (**R3**), "promising" (**R1**), "simple yet elegant" (**R2**) and "interesting" (**R2** , **R3** , **R4**). Reviewers appreciated that much prior work has involved hard-coding levels of linguistic structure into different model architectures (**R3**) and that our method is a "general plug-and-play model-agnostic method" (**R4**) that does not need to "decide on these points a priori" (**R3**).

Larger points: **R4** 's primary concern is the gap between the performance of our BERT + prism model and SOTA for the tasks we consider, and helpfully provides citations for SOTA models. However, the purpose of these experiments is to consider how tools from spectral analysis can affect the representations of a single, general-purpose model (BERT) across different tasks. This has both intrinsic scientific value as well as practical value, as these techniques could then be integrated into the various task-specific architectures. Two other factors explain this gap: 1) BERT contextual embeddings are known to perform surprisingly poorly on tasks beyond the word-level (see e.g. [2], where BERT embeddings perform up to 16 points worse than even GloVe embeddings) 2) SOTA models are trained end-to-end, while we train only a logistic regression on top of the token-level BERT embeddings without scale-specific pooling.

Other points: **R1** **: "The authors do not fully describe the various hypotheses they imply..."** This is a great point, and we'll make sure points 2(a), (b), and (c) that **R1** mentions are made explicit in the next revision.

R2 **: Prism layer transforms are fixed; how does this compare to a learned transform?** This is an interesting question. When pretraining with the prism layer, a fixed transform is useful because the MLM task is very local (Fig. 4) and the fixed bands encourage the model to produce multiscale representations. Learning the transform to optimize the MLM task may lead to all the frequency bands becoming local. However, it is an interesting direction for future work whether clever regularization can get the best of both worlds here.

R2 **: "In fig.5, why is BERT+prism worse for indices outside [200, 300]?"** We touch on this in L195-198, but the higher loss around the masked region suggests that the model relies more on the (redacted) distant context than BERT does to perform the MLM task, as we'd expect given the prism constraint.

R3 **: "precise choice on where to use the prism layer raises some questions..."** We agree that it would be very interesting to explore other ways of using the prism layer, including between every layer. We will include a comparison to this setting for the camera ready.

R3 **: "The way of dividing the embeddings into 5 sectors seems a bit naive"** We made this choice primarily to enable clear comparisons between tasks at different linguistic scales. In practice, one could choose particular scales relevant to desired end tasks or smoothly shift the frequency band between individual neurons as opposed to sectors of neurons. We will note this in the paper, and that there is opportunity for future work!

R4 **: "It would be nice to see ablations where you use high filters on POS tagging and low filters on paragraph/document classification to see the gains that come from choosing the right set of filters for each task."** We thank **R4** for the useful suggestion. For the lowest frequency sector of the BERT + prism model, topic classification accuracy is 45.1% (vs 51.0 for the full model and 5.3% for the highest-frequency sector). With the highest-frequency sector, POS tagging accuracy is 84.1% (vs 94.4 for the full model and 16.8% for the lowest-frequency sector). This suggests that the bands are largely but not entirely responsible for the high performance of BERT + prism, as expected.

R4 **: "As a sanity check, you could try to see what happens if you don't finetune the initial BERT model on wikitext-103?"** We thank **R4** for the useful suggestion. The original BERT model achieves an accuracy of 94.6% for POS tagging, 41.8 for dialog acts, 28.9 for topic classification, slightly worse than our model that was trained longer on WikiText-103 (95.9%, 47.1%, 32.2% respectively). As before, these are trained on contextualized token representations and are not comparable to models that perform pooling or are trained end-to-end. We will include this ablation in the next revision.

R4 **: "Since Figure 5 demonstrates good performance on long range masked language modeling, LAMBADA might be a good benchmark to validate this."** This is an interesting idea for future work, as it is not yet straightforward to use BERT for assigning probabilities to multi-token words in LAMBADA given 1) the constraints of the masked language modeling interface and 2) that LAMBADA lacks end of sentence punctuation needed for bidirectional conditioning.

Reviewers also offered a number of other suggestions which we are grateful for and will incorporate into the final version of our paper.

References: [1] Understanding intermediate layers using linear classifier probes [2] Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks