We thank all the reviewers for their thoughtful feedback. We first address common issues, and then individual comments.

R3 and R5 asked, respectively, (1) whether we are claiming that uniform quantization is *strictly better* than the other compression methods both empirically and theoretically, and (2) whether the strong performance of uniform quantization is surprising. In this work, we demonstrate empirically that a simple compression method, uniform quantization, can perform *similarly* to and sometimes outperform a variety of more complex methods. Theoretically, we show that we can lower bound the eigenspace overlap of uniformly quantized embeddings, which helps us understand the strong performance of uniform quantization at high compression rates. Importantly, in this work we are not attempting to demonstrate or prove that uniform quantization is strictly better than the other compression methods; as R3 correctly pointed out, there are times when the other compression methods perform as well (e.g., k-means throughout) or even slightly better (e.g., DCCL in lines 129-131) than uniform quantization. To us the strong empirical performance of uniform quantization is surprising because there has been recent work developing more complex compression methods (e.g., [2, 6, 18, 35]), and this work has not revealed that a simple method is competitive. We will clarify these points.

R2 and R3 had concerns about the amount of content we deferred to the appendix. We agree, and if accepted we will use the ninth page to include our logistic regression theorem, intuitions for our proofs, and more empirical results.

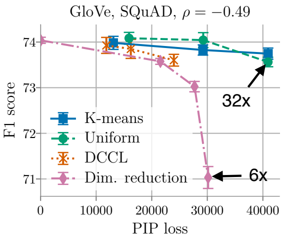
R2 and R5 suggested accelerating the computation of the eigenspace overlap score, for example by subsampling the vocabulary. We note that computing the eigenspace overlap between two 400k by 300-dimensional embeddings takes approximately 7 seconds on a 36-core Intel Xeon E5 CPU, and is thus already a fast operation relative to downstream model training. To further accelerate this, we run experiments with 1% vocabulary subsampling, and observe that across 10 random samples, the eigenspace overlap shows at most 4% relative change, and takes ~ 0.1 seconds to compute.

R2: R2 asked how we use embedding reconstruction error to evaluate compressed embeddings, given that reconstruction error cannot be applied to embeddings with different dimensions. R2 correctly pointed out this limitation of reconstruction error. In Appendix B.4, we discuss a variant of the embedding reconstruction error applicable to embeddings with different dimensions, where we consider the reconstruction error of the optimal projection of the lower-dimensional embedding into the higher-dimensional space.

R2 asked about our question answering results in Section 2.3. We use the DrQA model [5], as described in Section 4.

R3: R3 asked about the intuition for the proof of Theorem 2. We leverage the Davis-Kahan $\sin(\Theta)$ theorem, which upper bounds the amount the eigenvectors of a matrix can change after the matrix is perturbed, in terms of the amount of perturbation introduced. Because for uniform quantization we can exactly characterize the magnitude of the perturbation for any compression rate, this allows us to bound the change in eigenvectors for uniformly quantized embeddings.

R3 proposed an idea to use non-uniform quantization to further improve the performance of quantized embeddings. We are very excited to understand the impact of different quantization techniques as future work.



R4: R4 asked which points in Figure 1 correspond to the uniformly quantized (32X compression) vs. dimensionality-reduced embeddings (6X compression). We have updated the figure to clarify (see updated figure on the right).

R5: R5 asked whether the eigenspace overlap score can be seen as a natural generalization of the PIP loss [41]. While both the PIP loss and the eigenspace overlap score measure the quality of word embeddings, these metrics focus on different types of downstream tasks. In particular, while the PIP loss focuses on explaining the performance of embeddings on tasks which do not involve training a supervised model (e.g., word similarity), the eigenspace overlap score focuses on explaining performance on tasks which do involve training (e.g., sentiment analysis). From a theoretical perspective, in our work we bound the expected generalization error of downstream linear models in terms of the eigenspace overlap score; an analogous result has not been shown for the PIP loss. Empirically, we demonstrate that the eigenspace overlap score is more predictive of downstream performance than the PIP loss on a range of supervised tasks. We will clarify these points.

R5 asked about the differences between the proof of our generalization bound (Theorem 1), and the proof of PCA, which shows that the best rank-k approximation of a matrix is given by its top singular vectors. Although both proofs leverage the singular value decomposition, their objectives are different. While the goal of PCA is to minimize *reconstruction error* with respect to a single input matrix, the goal of our theorem is to characterize the difference in *generalization performance* between two embedding matrices.

R5 expressed concerns about the breadth of our empirical validation. Due to the space limit, we presented representative results in the main paper, and deferred the complete set of results to the supplementary materials (Appendix D). In our experiments, we compared the eigenspace overlap score with a range of baselines (PIP loss, Δ- and (Δ_1, Δ_2)-spectral approximation), and showed our results are consistent across a range of NLP tasks (sentiment analysis, question answering, GLUE tasks), embedding types (GloVe, fastText, BERT WordPiece), and compression methods (uniform quantization, k-means, DCCL, dim. reduction). We will transfer some of these results to the ninth page if accepted.