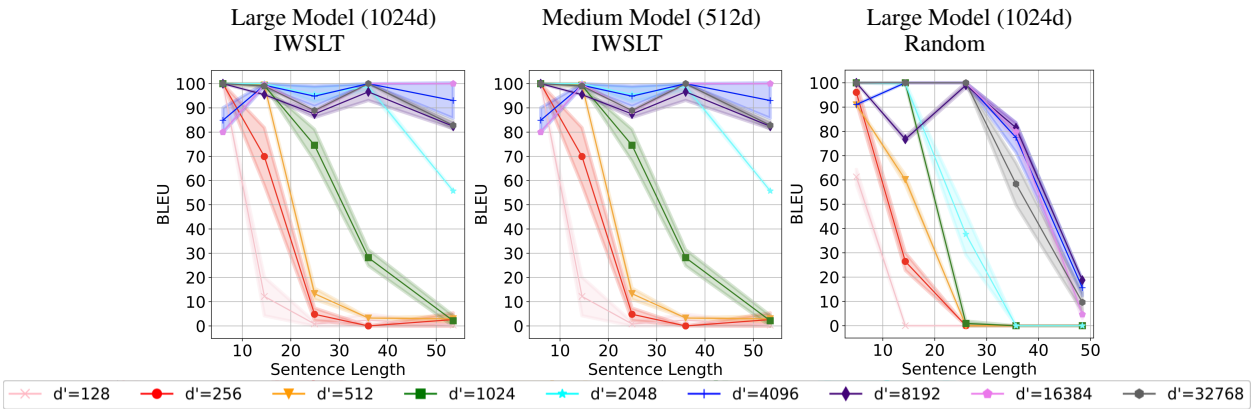
Author Response: We thank all our reviewers for their careful evaluation and numerous suggestions.

Reviewer 1 We agree that evaluating on just in-domain sentences is a limitation. As a result, we will add recoverability results on a subset of 50 sentences from the validation set of the IWSLT16 En-De translation dataset to the paper. This corpus is composed of TED talk transcripts, a very different style from the news text our language models were trained on. Further, we randomly sample another 50 sentences of varying lengths where each token is sampled randomly with equal probability with replacement from the vocabulary structure to evaluate whether z memorizes the sentence independent of the language model. We evaluate this across all our model sizes on the language models that have seen 50M sentences. For space constraints we provide two representative graphs over the IWSLT16 data (large and medium models) and one representative graph over the Random token sampling data (large model).



We are able to recover out-of-domain sentences well and thus seems to account for the domain shift from news to TED transcripts well. Recoverability estimates are nearly perfect for both the large and medium models. We observe that on the random data we are able to memorize sentences up to a length of approximately 25 with large z, but this drastically degrades after that–indicating that memorization does not fully explain results on Gigaword and IWSLT. We will add these to the paper. Our results and conclusions hold only for English, and this is a limitation of the work. We are also interested in your proposition about performing attention over smaller z's. It isn't equivalent to matrix multiplication and could accelerate learning, but we have to leave these for future work.

Reviewer 2 We missed the Mu et al 2017 paper on representing sentences with low-rank subspaces, which we have now added to the related work section. Our definition of the reparametrized sentence space is language-model agnostic and thus applicable to GPT-2. However, since GPT-2 uses a Transformer (not an RNN) decoder, we must change how we reparametrize the sentence space and figure out where to integrate z. One possible choice is to use z as an additional bias in each feed-forward layer of every word in a given sentence. To address the impact of optimization strategy, we reran a some of our best performing settings with Adam with a learning rate of 1e-4 and default parameters. We find that Adam results in z's that do not exceed 1.0 BLEU (Table 1). We will add this to the paper.

Reviewer 3 We don't explicitly assume a true length when decoding with beam search—we stop when an end of token or 100 tokens is reached. Some of the degenerate candidate recoveries repeat a specific token until the 100 token limit is reached. Larger beam sizes do not significantly change performance, see Table 2. We provide a partial table below and will add a couple of sentences to the paper stating this and clarifying the previous point.

Table 1: Optimizer results on English Gigaword

Model	$\|Z\|$	BLEU ConjGrad	BLEU Adam
SMALL; 50M	8192	81.1	0.031
MEDIUM; 50M	16384	92.4	0.234
LARGE; 50M	4096	99.8	0.037

Table 2: Beam width results on English Gigaword

Model	$\|Z\|$	BLEU Width=5	BLEU Width=10	BLEU Width=20
SMALL; 50M	512	40.0	40.3	40.5
SMALL; 50M	8192	81.1	79.8	79.6
MEDIUM; 50M	512	41.1	41.1	42.3
MEDIUM; 50M	16384	92.4	91.9	89.8
LARGE; 50M	512	54.8	54.1	53.8
LARGE; 50M	4096	99.8	99.8	99.5