We thank our reviewers for their time and valuable comments.

Motivation We have observed in the literature and also from personal communication at recent conferences incl. ICML and ICLR that almost all text GAN practitioners do not believe it is possible to train a GAN using REINFORCE on language with such high dimensional action spaces (e.g. vocabulary sizes of 10,000 or 20,000 as we have done). Instead the area publishes and promotes complex training techniques to overcome instability, using pre-training of the discriminator and generator; or periodic teacher forcing with a tuned schedule of regularity. Furthermore in some cases the GAN models are heavily pre-trained and only fine-tuned as a GAN with a miniscule learning rate (e.g. "Adversarial Feature Matching for Text Generation" Zhang et al. 2017).

We feel this paper will have a significant impact, by showing that stable training can be obtained with REINFORCE. We think this will re-focus the community from overcoming stability to benchmarking with richer data (EMNLP and WikiText are probably too small) and scaling with larger models — hopefully to a state where one can observe a significant difference in sample quality.

Overstatement The reviewers note that some of the wording in the paper suggests that we have solved GANs for text, which we agree is not the case. We do show ScratchGAN is producing samples of similar quality to language models for these datasets, however it is clearly a much worse generative model than the pure MLE variant and so far, the more compute-intensive GAN training is not providing us with a much better model. We will tone down any language which suggests ScratchGAN outperforms MLE. But we stand by the observation that sample quality and diversity appears to be close to the MLE model, and some metrics confirm this (BLEU, FED).

Sample quality We certainly agree that neither the MLE or ScratchGAN are producing groundbreaking samples - and we mostly attribute this to choice of dataset. We ran this on two LM datasets that had been benchmarked by prior GAN work (EMNLP and WikiText). This had the benefit of comparison to prior work for objective measures BLEU/self-BLEU (Fig 2a). However it has the downside that the samples are quite bad and make for a difficult qualitative comparison. We think future work should focus on scaling to larger datasets, generating larger bodies of text and using aggregate human evaluation to provide a more objective sense of sample quality.

Reviewer 1 We agree with your point that we are dismissing non-autoregressive language models. We will add a couple of sentences to highlight progress in feed-forward approaches.

We have addressed these typos, thank you for noting them! Also, we have increased the tables 7 and 9 from 5 samples per (model, dataset) to 15.

Reviewer 2 Since there is a symbiosis in ScratchGAN between the discriminator and generator — both are recurrent models over text, we do believe a promising direction would be to scale the dataset and model (e.g. to a transformerxl) to obtain better quality samples.

We agree MLEs are still superior to GANs, we have not solved GANs for text but we think this work is an important data point along the path to doing so. Please see our *Overstatement* section — in short we will tone down the claims of the paper. Re. code release, we are in the process of trying to release a simple colab script for training, such that people can see all of the components working.

Reviewer 3 We agree the sample quality is not very good, we partly address this in the *Sample quality* section above. It was not clear to us that there was a qualitative difference between MLE and ScratchGAN, some MLE samples are quite degenerative, e.g. "*after the sets of UNK wear UNK and UNK ' UNK ' UNK to tell him , UNK UNK they play UNK UNK with UNK around a UNK* ." However future work should benchmark these approaches at scale, with a larger model and use a cumulative human evaluation to assess qualitative appearance; alongside the automatic scores.

For objective measures, we do compare to existing GAN approaches (Fig 2a). However the real objective is to have GANs considerably outperform MLE, since it is agreed this is still the best approach for text generation.

Re. overclaiming, we agree ScratchGAN does not outperform MLE and have toned down any language that appears to make this claim (see *Overclaim* section above). We genuinely do not want to claim that ScratchGAN solves GANs for text, just that it is possible to train a GAN to a decent level of quality (judged by objective measures) without a complex training procedure of pre-training, teacher forcing, Gumbel Softmax with a scheduled temeprature increase etc.

Good point re. pre-trained word embeddings. We decided to keep them because there was no change in performance within the ablation study — for this comparative run the model was truly from-scratch ;-).

Aside from language applications, we think the result of this study — that REINFORCE can be stably trained in this challenging setting — will be of interest also to reinforcement learning practitioners that are interested in high-dimensional action spaces; e.g. for medical treatment prediction in electronic health record time-series. Thus we ask you to consider this core research contribution, when reconsidering your score.