Thanks to all the reviewers for the insightful comments and feedback.

- About the use of pretraining (R1,R2,R3,R4) Our text GAN is the first to outperform MLE, to the best of our knowledge (based on our results and those from the ICLR'20 paper "Language GANs falling short"). While it also strongly outperforms previous GANs, we are the first to leverage on self-supervision pretraining (BART/T5). Hence, all reviewers pointed out that comparison is unfair, and wonder how much comes from the pretrained language model. First, please note that **most previous works do use MLE pretraining as us**, ScratchGAN is the exception. And **none of them report results that outperform their corresponding MLE** (including ScratchGAN), as opposed to us. Also, to further analyse the results, we complete Fig. 2 from the paper with the performance of 2 additional models (Fig. 1 below, under *rebuttal*). We observe that i) when initialised with T5, ScratchGAN under-performs MLE, as opposed to our proposed ColdGAN; and ii) when randomly initialised, our ColdGAN is the only GAN to outperform MLE.

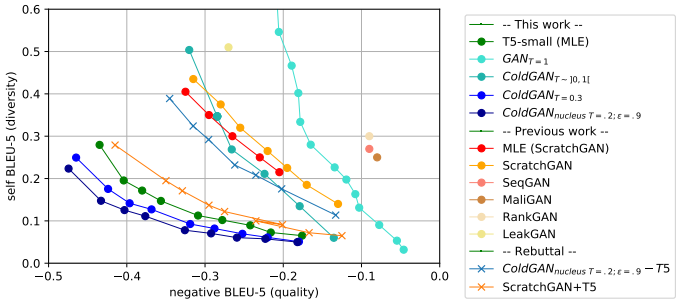

Figure 1: Results on the EMNLP 2017 News dataset (for all metrics, lower is better). Scores for previous works are taken from "Training language gans from scratch". *Scratch GAN+T5*: Scratch GAN but this time initialised with pretrained weights from T5; *ColdGAN-T5*: ColdGAN, but not initialised with T5 pretrain;

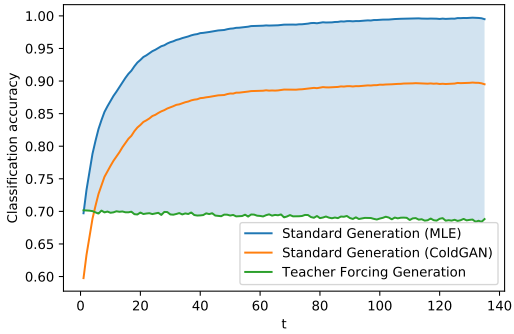

Figure 2: We completed the figure with our GAN outputs in *Standard* generation mode, as requested by R4. The discriminator is less able to distinguish than when generation comes from MLE.

- About finetuning (R1) There is a misunderstanding here, that we feel had a strong impact in the evaluation: we agree that without the finetuning of baselines, our experiments would be misleading. But all our experiments actually considered models finetuned via MLE to the specific task. While it was not explicitly written in the paper, this was somehow contained in the "MLE" name of our baselines. Please note that implementation details are given in appendix.

- Regarding Fig. 1 (R1,R2,R3,R4) In Fig. 1 from the paper, we used our baseline T5-small. We will make this explicit if accepted. To answer R1, the objective of this figure was 1) to compare it with the standard generation to highlight the exposure bias and 2) to assess the ability of the discriminator on these two modes. There is clearly an ability to distinguish human-generated from teacher-forced generation since the accuracy is higher than 50%. To answer R2, accuracy starts at 70 %, indicating that even with a very short prefix the discriminator is good at distinguishing real from generated texts. To answer R3 and R4 (who have opposite feelings w.r.t. the reported results), please note that results are consistent with previous works as written line 24, including those of the "Real or Fake?" paper suggested by R4 (70% accuracy for T=1 corresponds to the score this paper reports for its smaller Relative prefix length). The good accuracy of discriminators, even for short prefix, can be explained by the fact that they are trained on the output distributions of the corresponding generators. This specialization allows them to well distinguish human from generated texts for these specific generation distributions (but not for every modified distributions as shown in table 1 of the paper). In Fig.2 above, we show that discriminators have eventually more difficulties to distinguish real from generated texts with a generator resulting from our ColdGAN approach than with the MLE baseline.

- About Improvements (R3) The improvement for automatic metrics is not large, but i) it holds true for all of them and across all the three tasks; ii) when changing over 10 different seed initialization, the variance of the improvement is clearly above the last SOTA, as written line 259. Note that improvement is rarely in a large magnitude for these highly competitive tasks; iii) the automatic metrics are arguably not perfect, which is the reason why we conducted a human evaluation, showing that our model significantly outperforms MLE for the fluency (see table 3). For fluency (question 2 of R3), we used the following definition: sentences perceived by a human as natural and grammatically correct.

- About Novelty (R3) Through our preliminary analysis, we are the first to show that classic discriminators in text GANs are not stable around the generator distribution mode, which prevents them to outperform MLE performances. We thus propose a new training methodology, based on importance sampling to focus on areas of best interest of the representation space, that i) is not biased and ii) succeeds at stabilizing the discriminator, hence the reward, improving thus significantly the entire training dynamic.