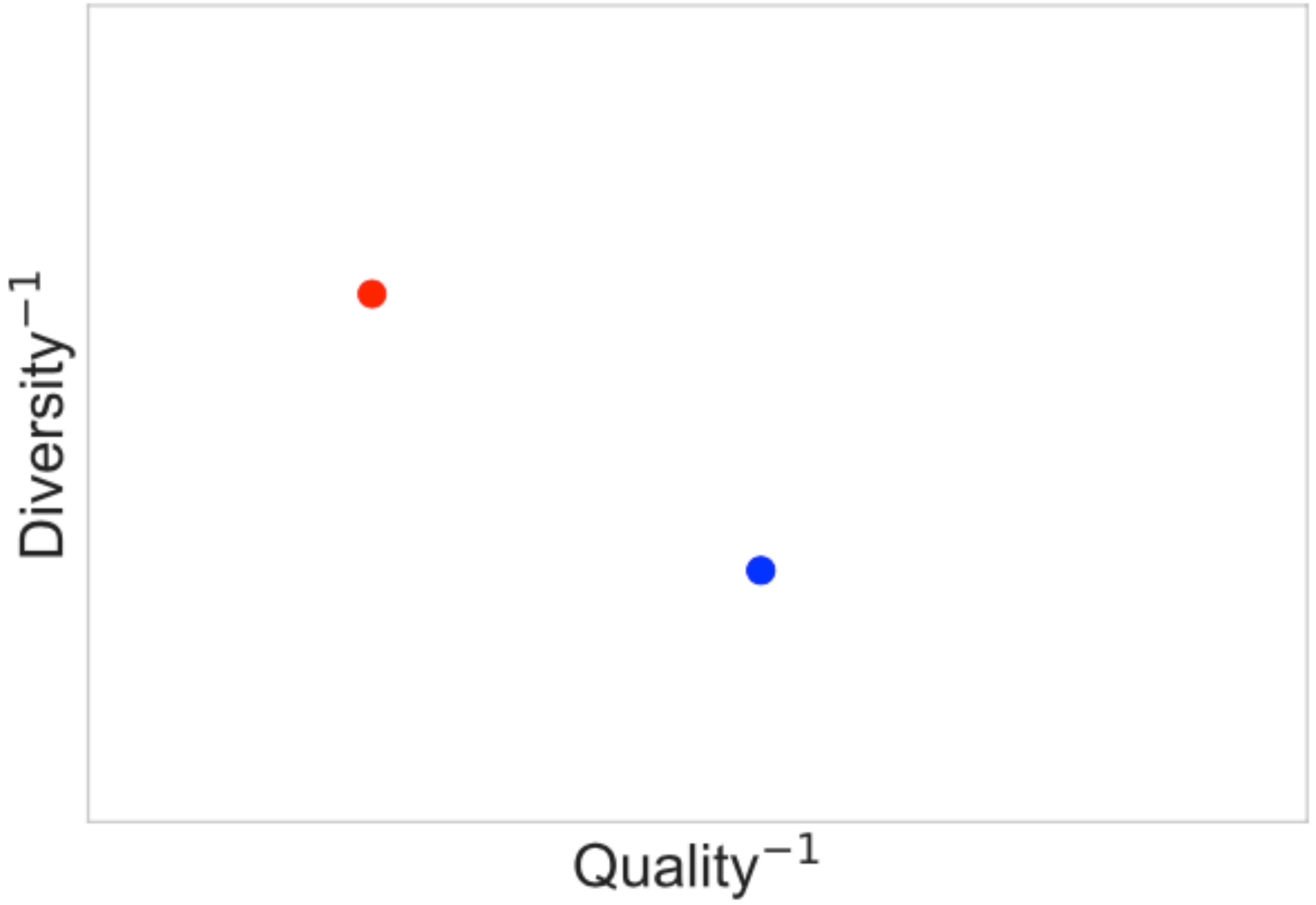
LANGUAGE GANS FALLING SHORT

Морозов Дмитрий, БПМИ172

Natural Language Generation — оценка качества

- Точность (грамотность, логичность и т.д.) каждого сгенерированного предложения
- Разнообразие (diversity) каждого сгенерированного предложения



Adversarial Text Generation

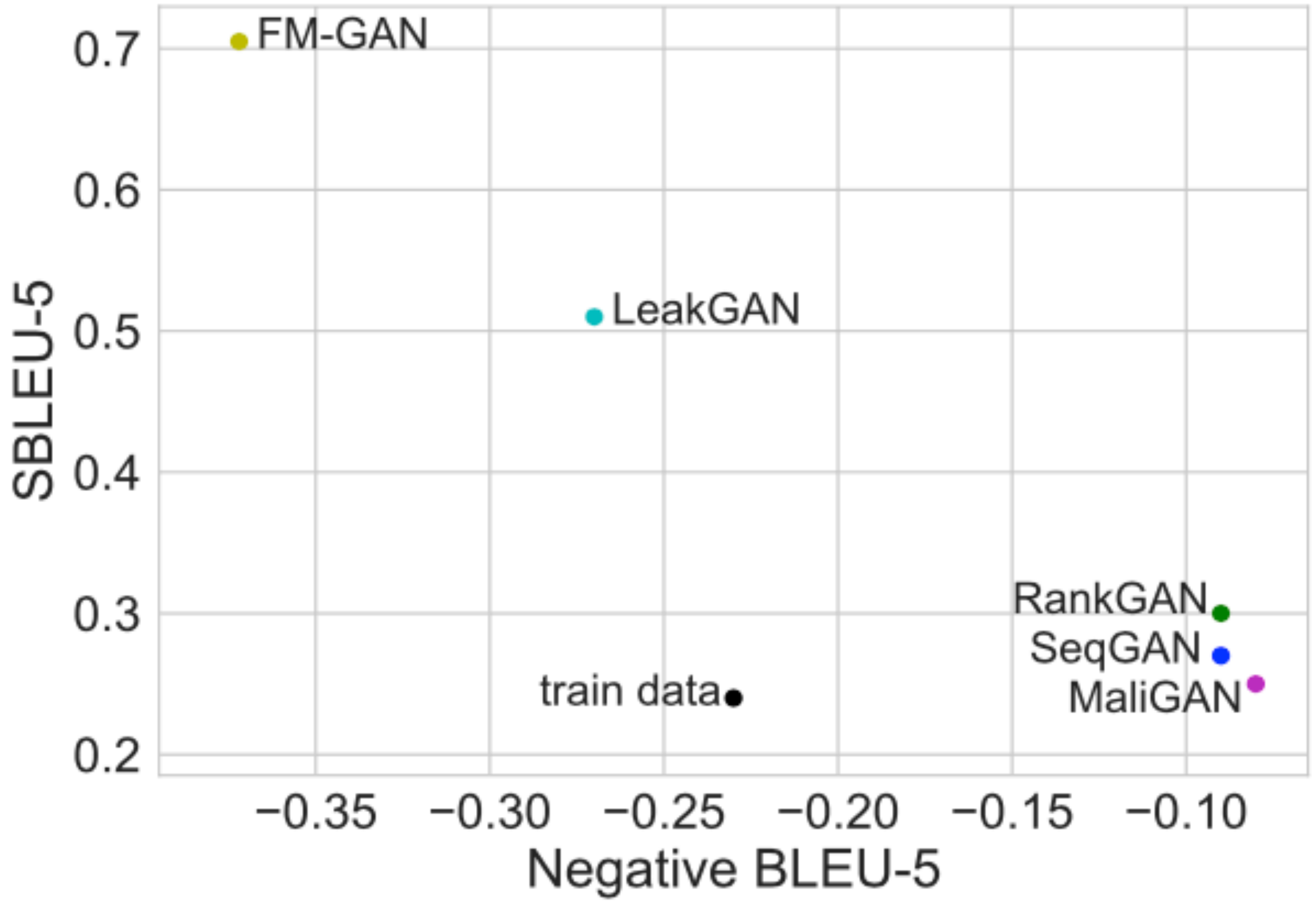


Figure 2: Negative BLEU-5 versus SBLEU-5 *(lower is better for both metrics)* on the EMNLP2017 News dataset taken from (Lu et al., 2018b) and this work (train data and FM-GAN). These scatter plots do not clearly show which algorithm is preferred since none strictly dominates on both metrics simultaneously.

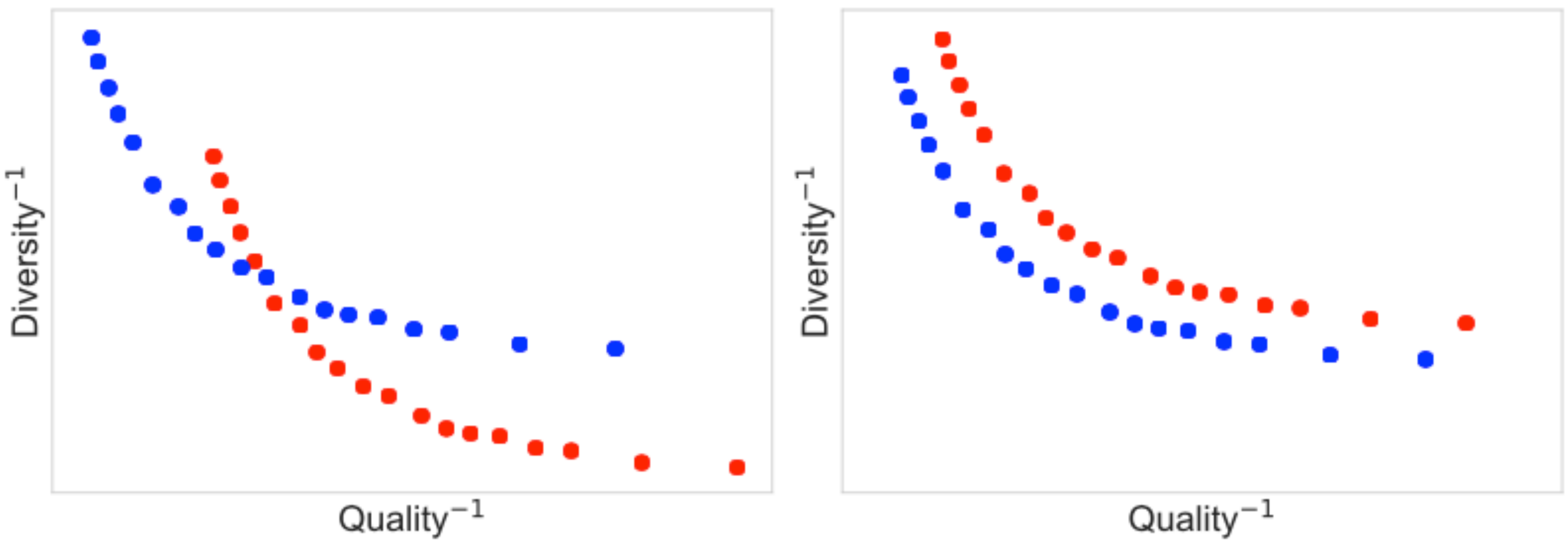
Natural Language Generation — оценка качества

- Temperature Sweep — использование температуры в качестве совместной оценки качества и разнообразия

$$G_\theta(x_t \mid x_{1:t-1}) = \text{softmax}(o_t \cdot W/\alpha)$$

α	Samples
2.0	(1) If you go at watch crucial characters putting awareness in Washington , forget there are now unique developments organized personally then why charge . (2) Front wants zero house blood number places than above spin 5 provide school projects which youth particularly teenager temporary dollars plenty of investors enjoy headed Japan about if federal assets own , at 41 .
1.0	(1) Researchers are expected to comment on where a scheme is sold , but it is no longer this big name at this point . (2) We know you ’ re going to build the kind of home you ’ re going to be expecting it can give us a better understanding of what ground test we ’ re on this year , he explained .
0.7	(1) The other witnesses are believed to have been injured , the police said in a statement , adding that there was no immediate threat to any other witnesses . (2) The company ’ s net income fell to 5 . 29 billion , or 2 cents per share , on the same period last year .
0.0	(1) The company ’ s shares rose 1 . 5 percent to 1 . 81 percent , the highest since the end of the year . (2) The company ’ s shares rose 1 . 5 percent to 1 . 81 percent , the highest since the end of the year .

Temperature Sweep



GAN vs MLE — SYNTHETIC DATA EXPERIMENT

Model	NLL_{oracle}
SeqGAN (Yu et al., 2017)	8.74
RankGAN (Lin et al., 2017)	8.25
LeakGAN (Guo et al., 2017)	7.04
IRL (Shi et al., 2018)	6.91
MLE ($\alpha = 1.0$)	9.40
MLE ($\alpha = 0.4$)	5.50
MLE ($\alpha = 0.001$)	**4.58**

Table 2: NLL_{oracle} measured on the synthetic task *(lower is better)*. All results are taken from their respective papers. An MLE-trained model with reduced temperature easily improves upon these GAN variants, producing the highest quality sample.

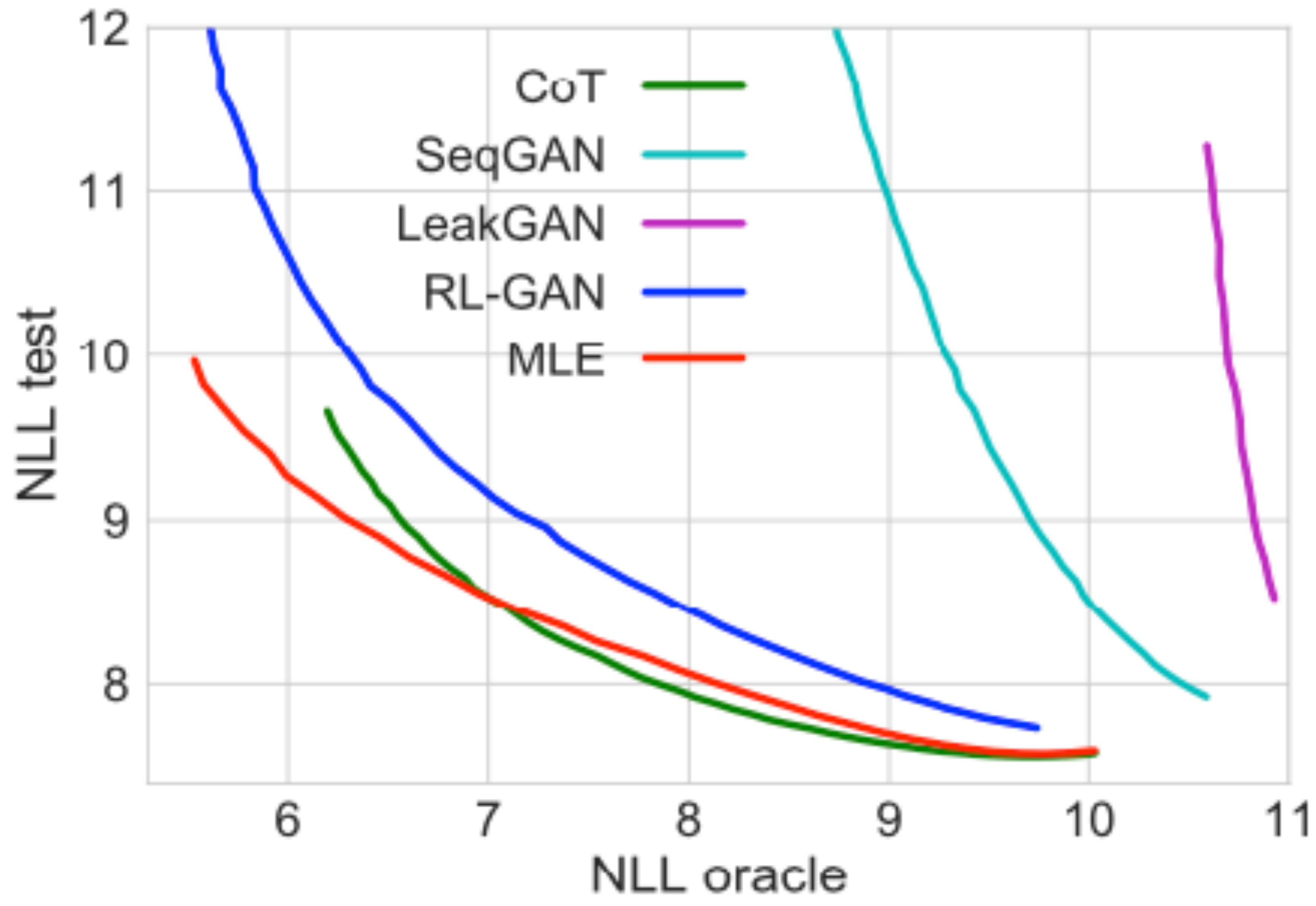
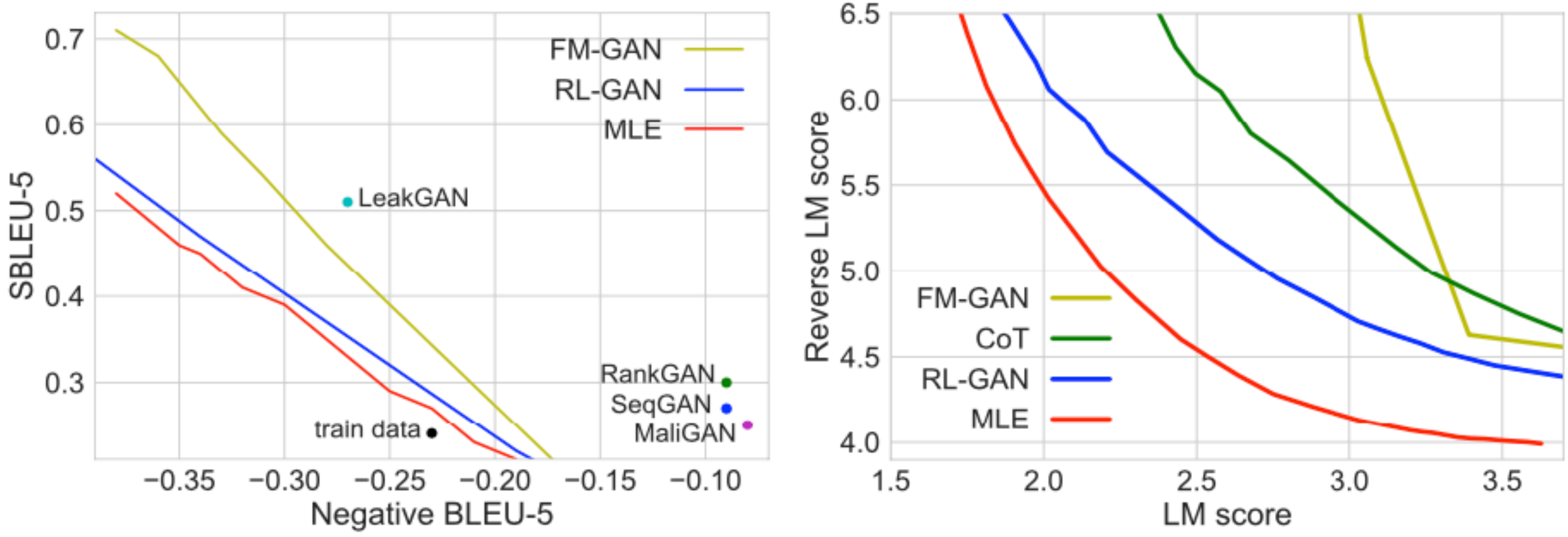


Figure 3: Effect of temperature tuning on the global metrics *(lower is better for both metrics)* for the synthetic task.

GAN vs MLE — LONG-TEXT GENERATION

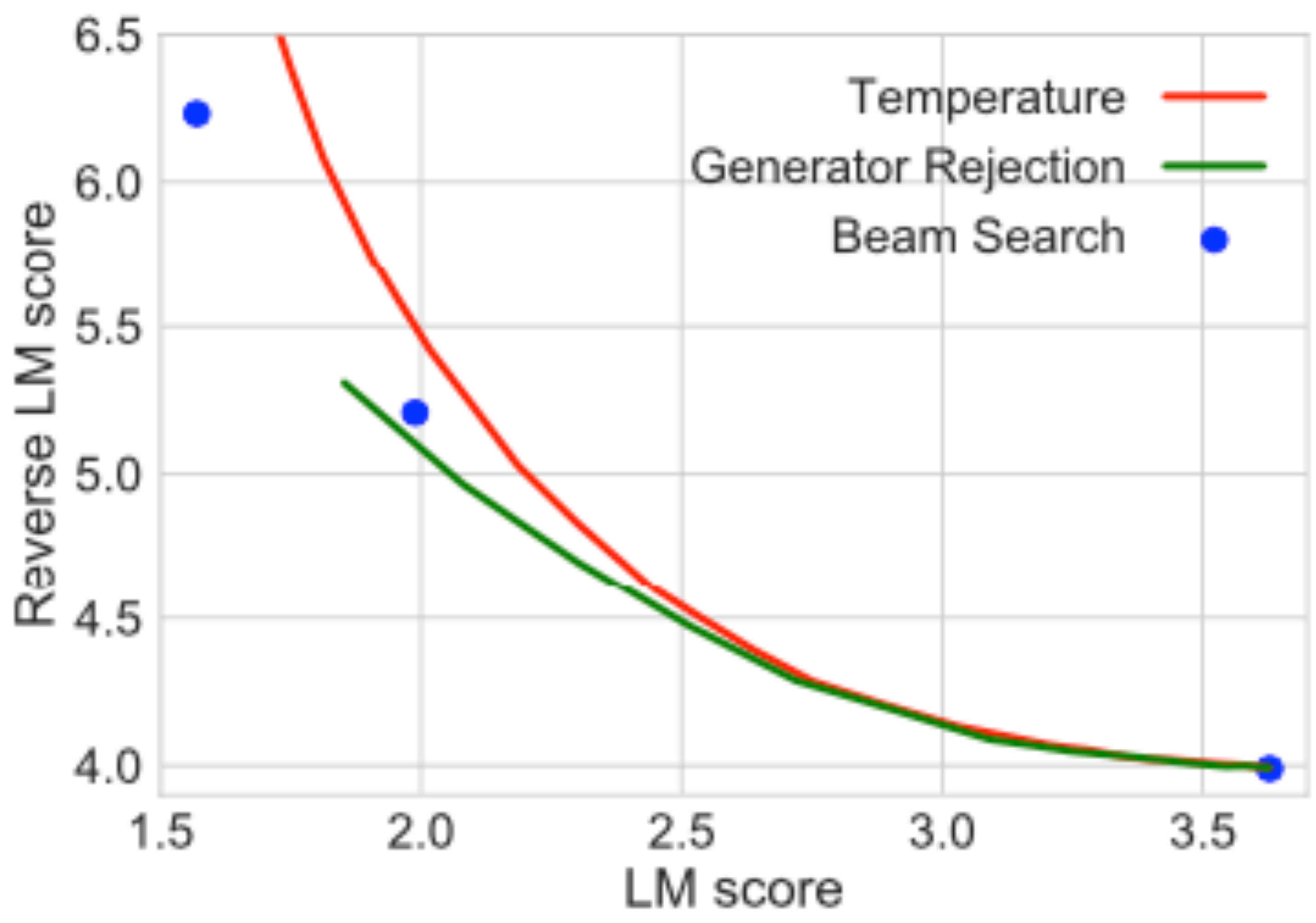


(a) Local metrics

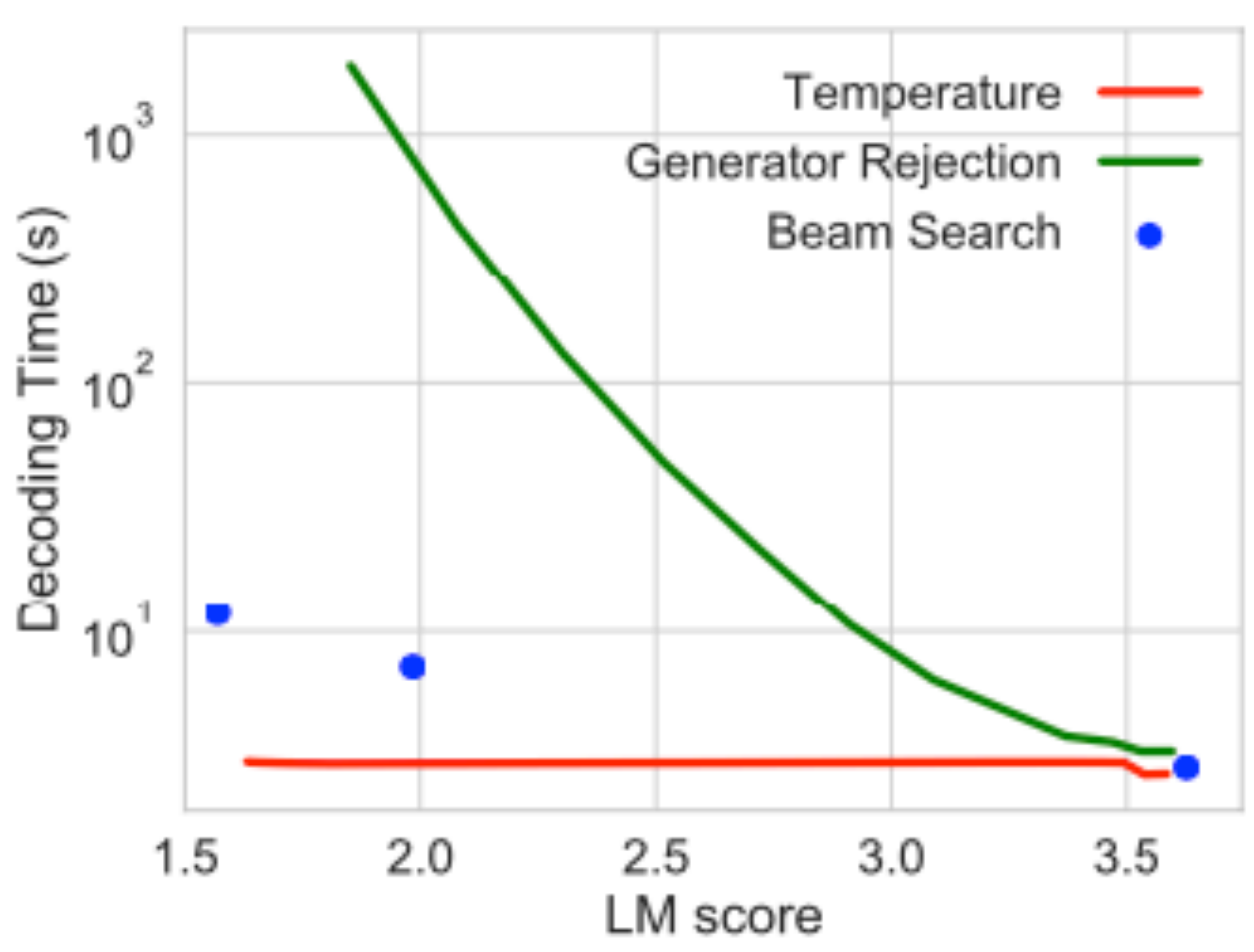
(b) Global metrics

Figure 4: Results on the EMNLP 2017 News dataset. *(lower is better for all metrics)*. MLE under a temperature sweep achieves better quality-diversity trade-off compared to the GAN approaches.

Temperature vs Beam Search vs Generator Rejection



(a) Diversity and quality



(b) Decoding time and quality

Figure 6: Analysis of decoding methods. *(lower is better for all metrics)*. **Left:** Less biased methods provided a better quality/diversity trade-off. **Right:** However, they are computationally much more expensive.

Список источников

- https://arxiv.org/abs/1811.02549

Список вопросов

1. Опишите, какой подход авторы статьи предлагают для совместной оценки качества и разнообразия генераций текстов. Запишите формулу условного распределения генератора при использовании этого подхода.

2. Что происходит с генерациями при увеличении/уменьшении температурного параметра Больцмана в методе Temperature Sweep с точки зрения согласованности и разнообразия?

3. Какие методы декодирования, согласно результатам экспериментов авторов, позволяют получить лучшие результаты, чем Temperature Sweep? В чем при этом их недостатки?