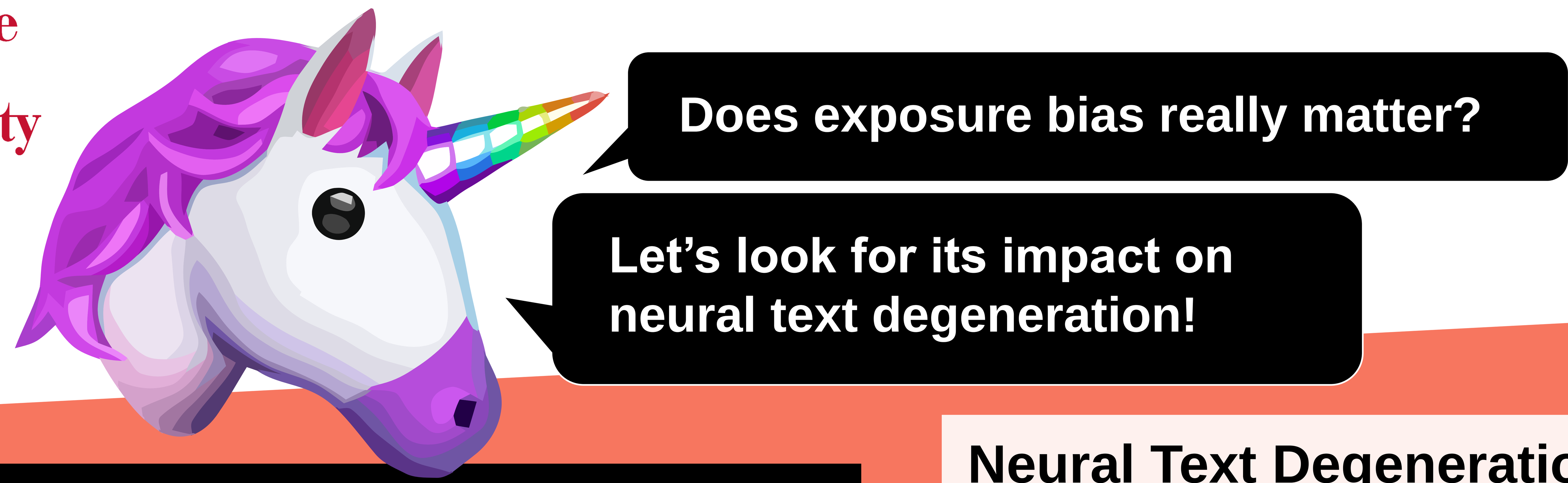
Carnegie Mellon University

National Taiwan University 國立臺灣大學

MIULAB



Relating Neural Text Degeneration to Exposure Bias

Ting-Rui Chiang, Yun-Nung (Vivian) Chen

Mistakes are made in the early phase

1. Qualitatively: classify real / artificial.

condition | generation ✗

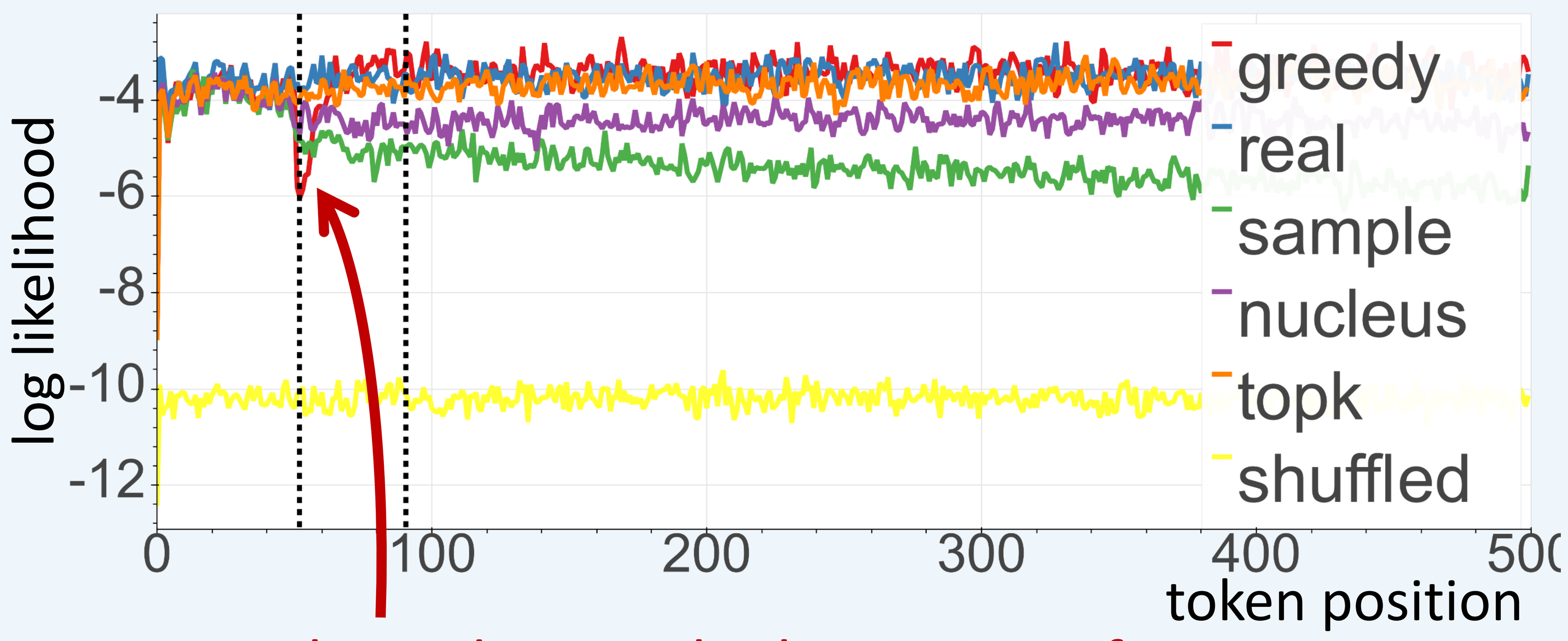
condition | real ✓

Human accuracy **78%**

condition | generation | loop | loop | loop

Implication: This part is unnatural.

2. Quantitatively: estimate likelihood with MLM

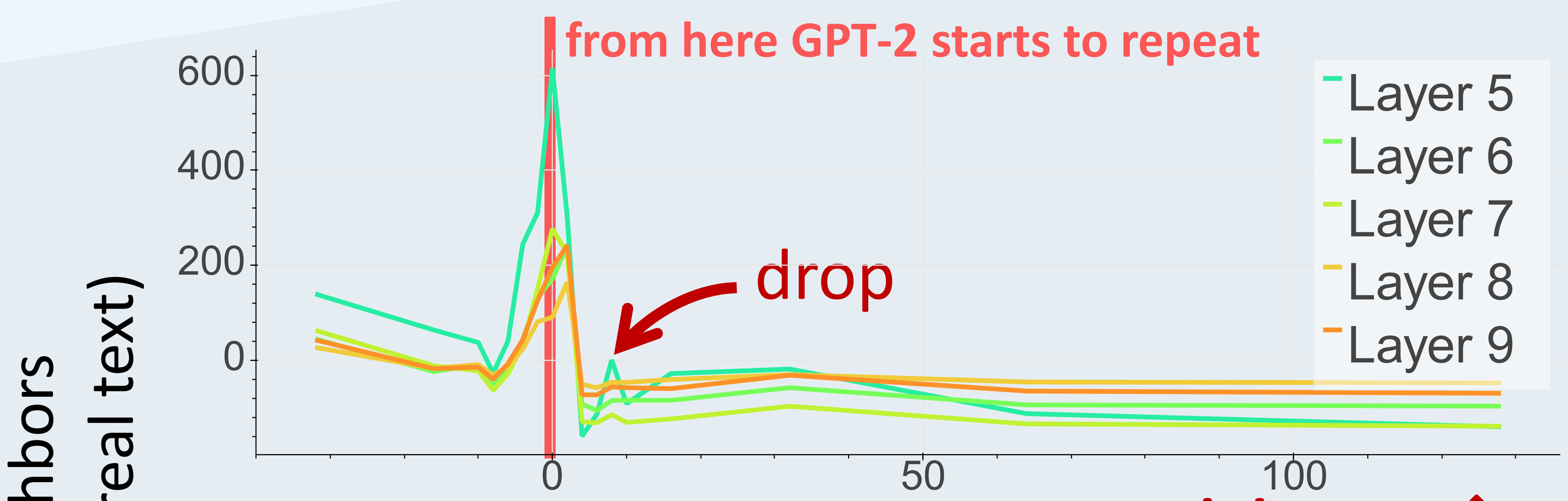


A sharp drop at the beginning of generation.

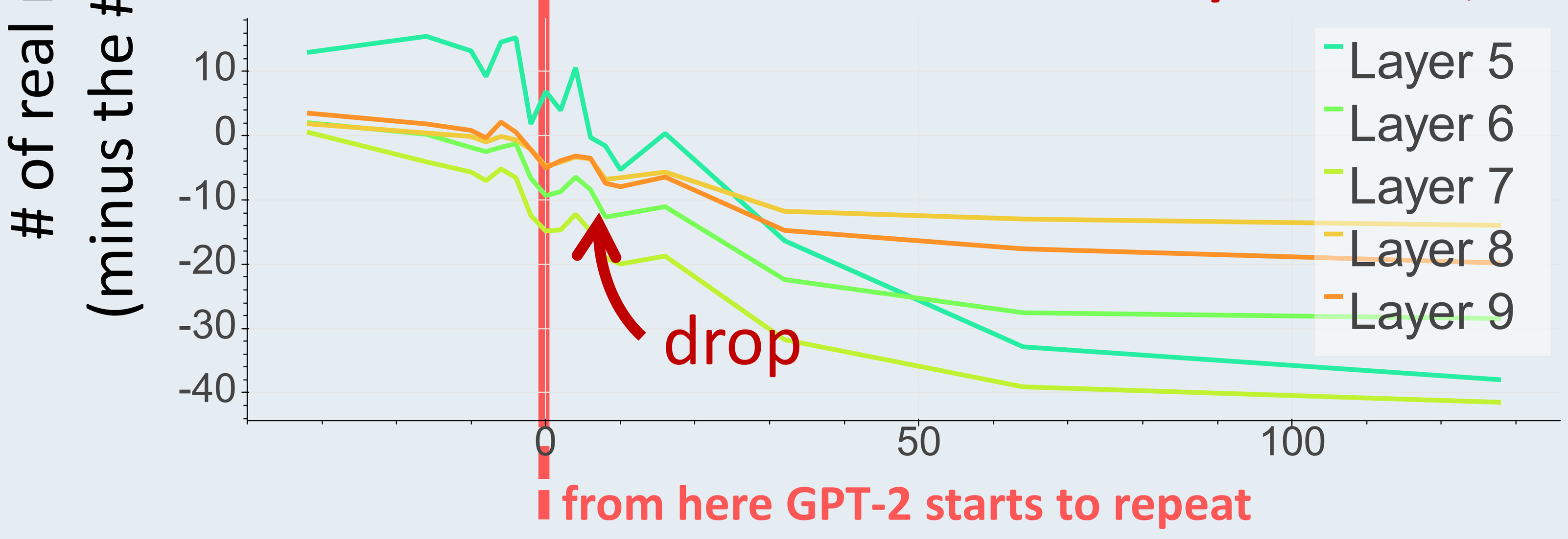
Indication: Mistakes are made at the beginning.

A Closer Look at the # of Neighbors

x-axis: token position relative to $\rho + \lambda$



Real neighbors are from GPT-2's **training** set ↑ **dev/test** set ↓



Neural Text Degeneration

With greedy decoding, GPT-2 just repeats…

We first saw Anki Overdrive, the company's follow-up to the original game, in the early 2000s. It was a game that was a bit of a hit, and it was a game **that was a bit of a hit** that was a bit of a hit that was a bit of a hit that …

ρ $\quad \rho + \lambda \quad \rho + 2\lambda$

condition | generation | loop | loop | loop | …

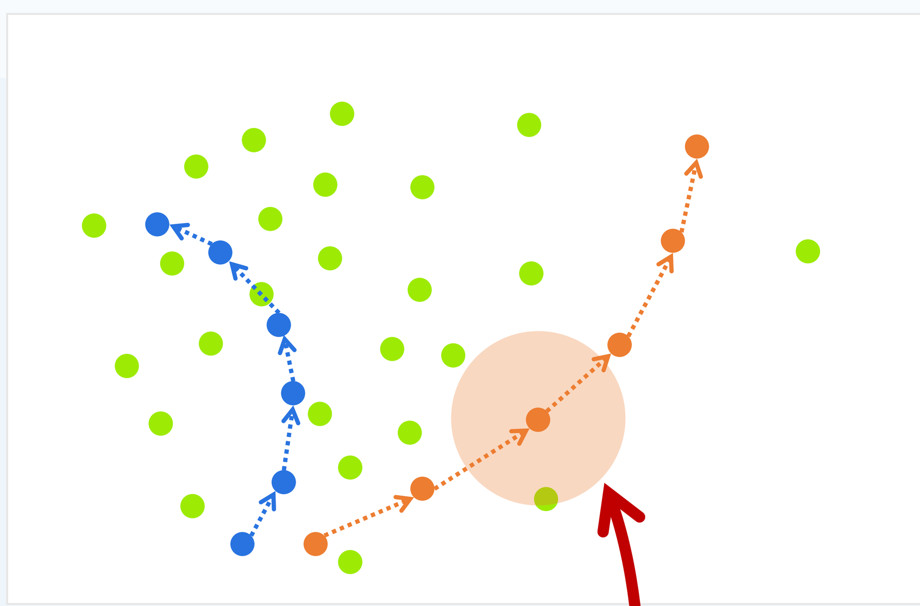
$\lambda \quad \lambda \quad \lambda$

Indications of Exposure Bias

1. Mistakes are made in the early phase
   - Qualitatively: unnatural text.
   - Quantitively: low likelihood.
2. Mistakes are significant to the model
   - lead the model to a state unseen in training time

Mistakes are significant to the model

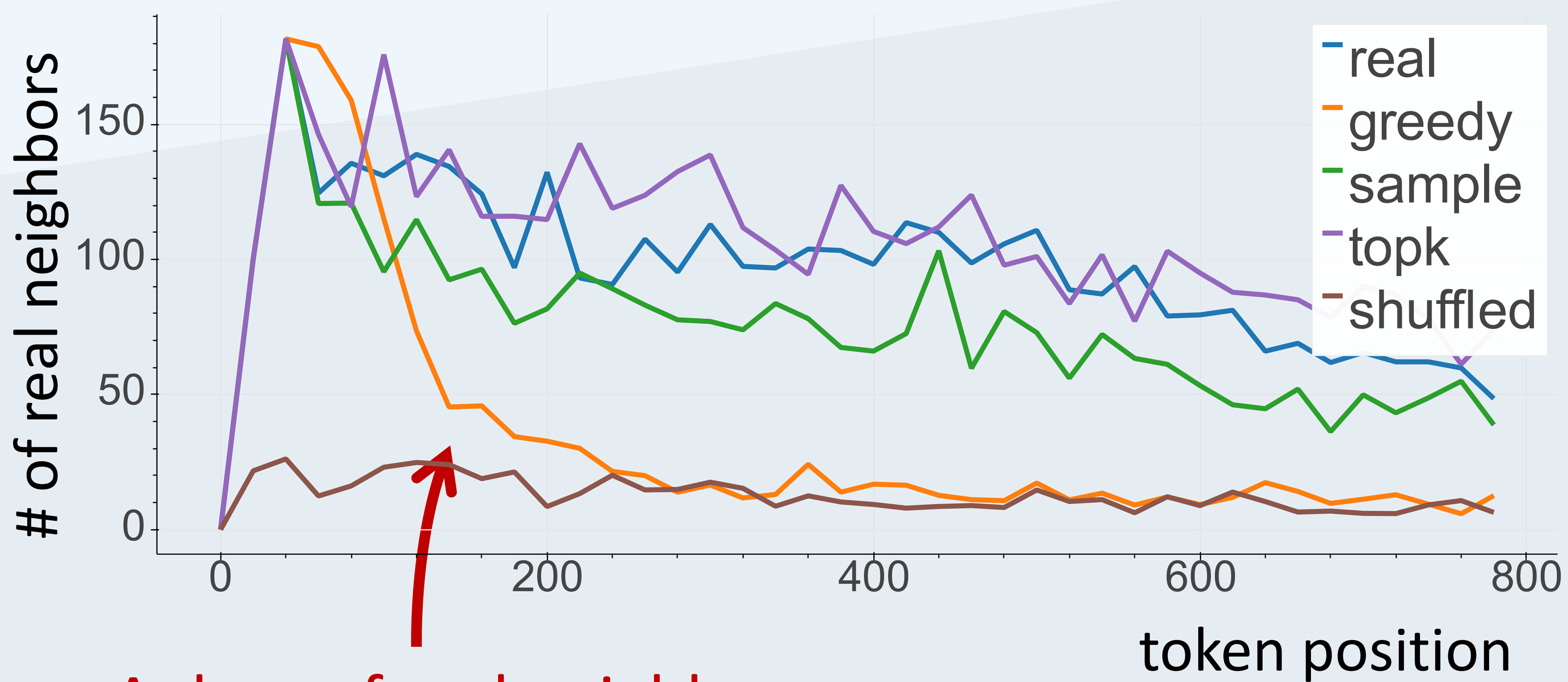
Compare the states to real states:



Real states: generated by encoding real text.

Artificial states: states of a generated sentence.

Count the number of real neighbors.



A drop of real neighbors.

Indication: Mistakes are significant to the model.

Takeaways

1. We summarize the indications of exposure bias.
2. We design the associated experiments.
3. Our results indicate the relation between exposure bias and neural text degeneration.