Mirostat:
A Neural Text Decoding Algorithm that Directly Controls Perplexity

Sourya Basu
University of Illinois at Urbana-Champaign

Govardana Sachitanandam Ramachandran
Salesforce Research

Nitish Shirish Keskar
Salesforce Research

Lav R. Varshney
Salesforce Research
University of Illinois at Urbana-Champaign

Neural Text Decoding

Language Modeling

- Language modeling is an unsupervised learning task of learning the probability distribution $p(x)$ from a set of examples of the form $x = (x_1, \ldots, x_n)$ where each $x_i \in \mathcal{V}$ and $\mathcal{V}$ is a finite set denoting vocabulary.
- $p(x) = \prod_{i=1}^{n} p(x_i \mid x_{<i})$
- Current state-of-the-art methods train a model with parameter θ minimizing the loss function

$$\mathcal{L}(T) = -\sum_{k=1}^{|T|} \sum_{i=1}^{n} \log p_\theta(x_i^k \mid x_{<i}^k),$$ over dataset $T = \{x^1, \ldots, x^{|T|}\}$.

- A trained model p_θ can be used for generating the ith word from the previous words by sampling from the distribution $p_\theta(x_i \mid x_{<i})$.

A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," *Unpublished manuscript*, Feb. 2019. [Online]

Z. Dai, Z. Yang, Y. Yang, J. G. Carbonell, Q. Le, and R. Salakhutdinov, "Transformer-XL: Attentive language models beyond a fixed-length context," in *Proc. Assoc. Comput. Linguist. Annu. Meet. (ACL 2019)*, Jul. 2019, pp. 2978–2988.

Top-k and Top-p (nucleus) sampling

- Trained model p_θ often contains unreliable tail distribution, hence several methods have been considered to truncate this tail.
- Top-k sampling generate texts by sampling from the top k most probable words/tokens. Here k is chosen in an ad-hoc manner to generate good-quality texts.
- Top-p sampling truncates the low probability tail of p_θ by sampling from $k(p)$ most probable words such that the cumulative distribution of these $k(p)$ words sum up to p.
- In a way top-p sampling dynamically changes k for each sampled token keeping p as a constant.

Repetitions and Incoherence

Examples: Repetitions

Observed average surprise value = 1.471 ***top-p sampling*** ***p=0.4***

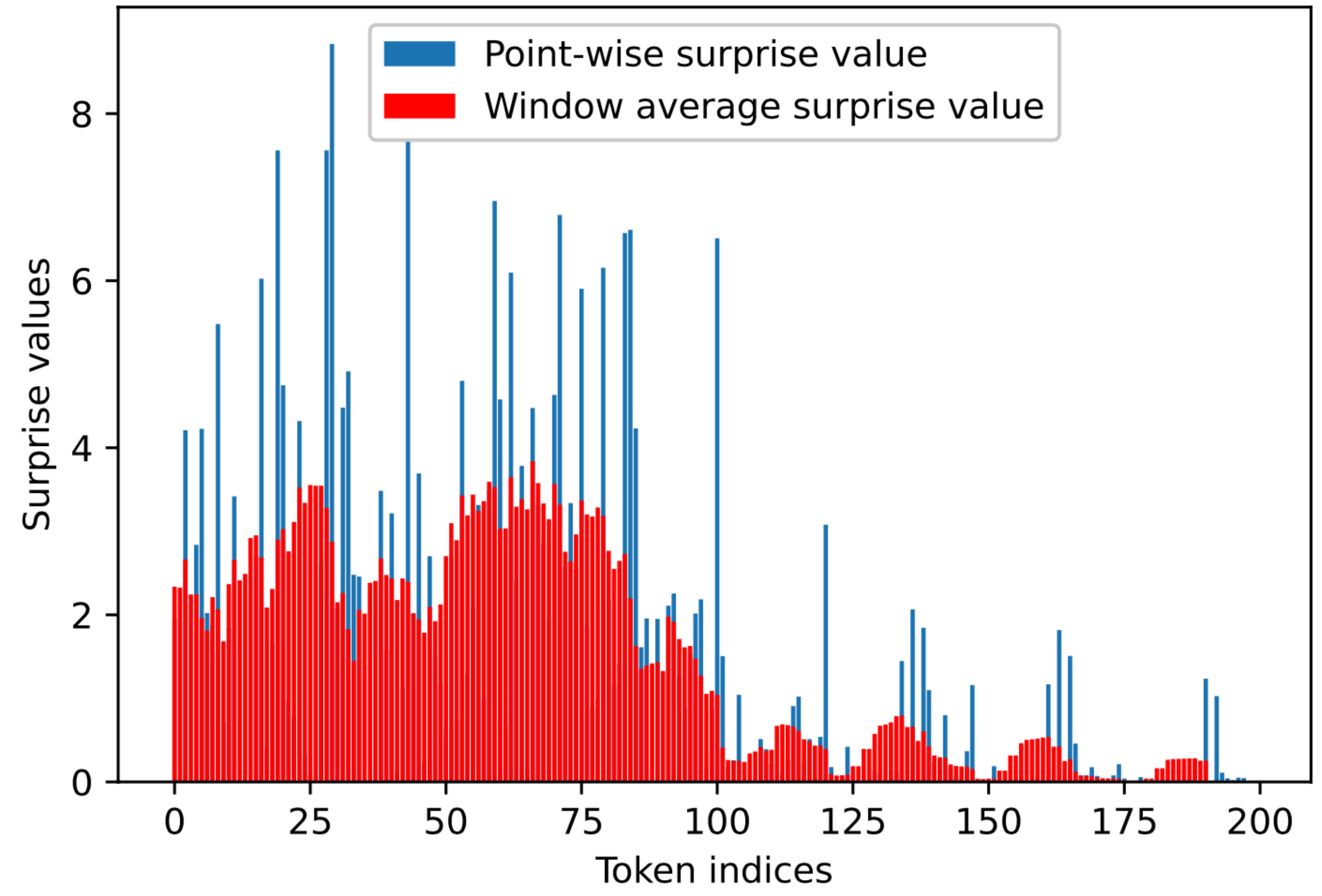
Generated text *= "Turing's work on the cryptography of the Kriegsmarine was based on the cipher used by the Kriegsmarine's famous E-2. The Kriegsmarine code was based on the Cryptographic Algorithm (CAT) of the Royal Navy. It was a new method of solving the problem of encryption, which was a mathematical problem which was the basis of the cryptanalysis of the Japanese naval cipher.*

Turing's work on the cipher was based on the algorithm used by the Kriegsmarine's famous E-2. The cipher was based on the algorithm used by the Kriegsmarine's famous E-2.

Turing's work on the cipher was based on the algorithm used by the Kriegsmarine's famous E-2.

Turing's work on the cipher was based on the algorithm used by the Kriegsmarine's famous E-2.

Turing's work on the cipher was based on"



(a) Top-p sampling with $p = 0.4$ and average observed surprise = 1.471

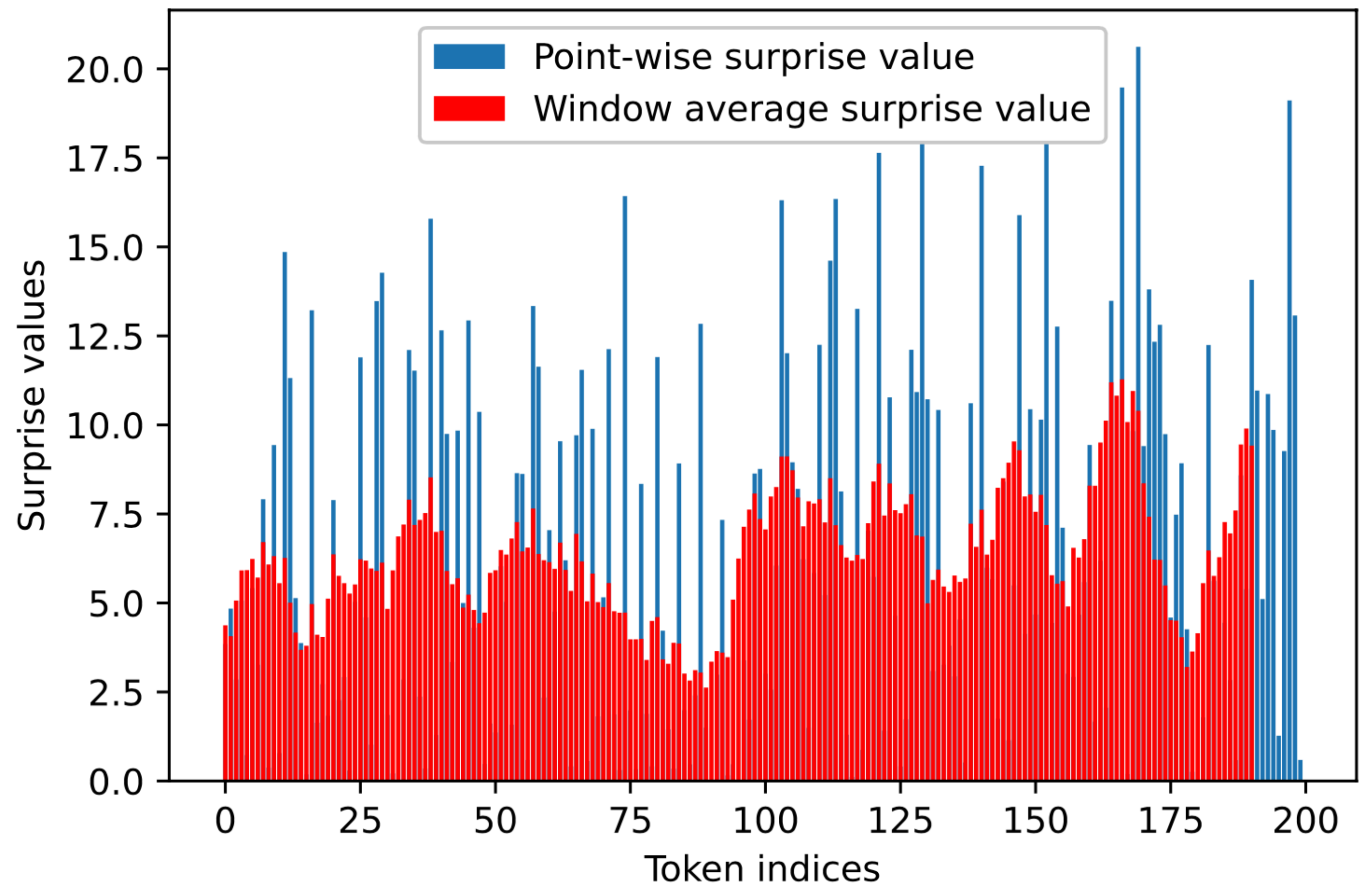
Examples: Incoherence

Observed average surprise value = 6.342 ***top-p sampling*** ***p=1.0***

Generated text = *"In January of 1942, in an open letter delivered to 8 persons and addressed to Sherwood, Shannon took the opportunity to address skeptics about the little out-of-print $128,000 unreported race as "cryptanalysis apparatus of extraterrestrial intelligence." His reply was also reprinted in 6 South Carolina publications (Atlanta Press, Fall Publishing, Miami, New House Press, Yorkshire Press, Sydney Press, Weatherby Press, Winter Press, and Thames & Hudson, to name a few), and eventually accepted as the final subscribed FHAT article. (Summary, were whom)*

Birmingham City Police concludes that [Turing's][new replace Robert] Symington is not guilty of breach of OKF law. Further, his suggestions for preventing his sentencing Should he reside anywhere near the city he defies extradition to Bavaria closer to dawn minimal? Lylimud Simpson (Ladies & Gentlemen, Lawrence Livermore University Press, Philadelphia): Jim Gilmore and its wife, Eva Civilsky,"

•



(c) Top-p sampling with $p = 1.0$ and average observed surprise = 6.432

Analysis of Top-k and Top-p (nucleus) decoding

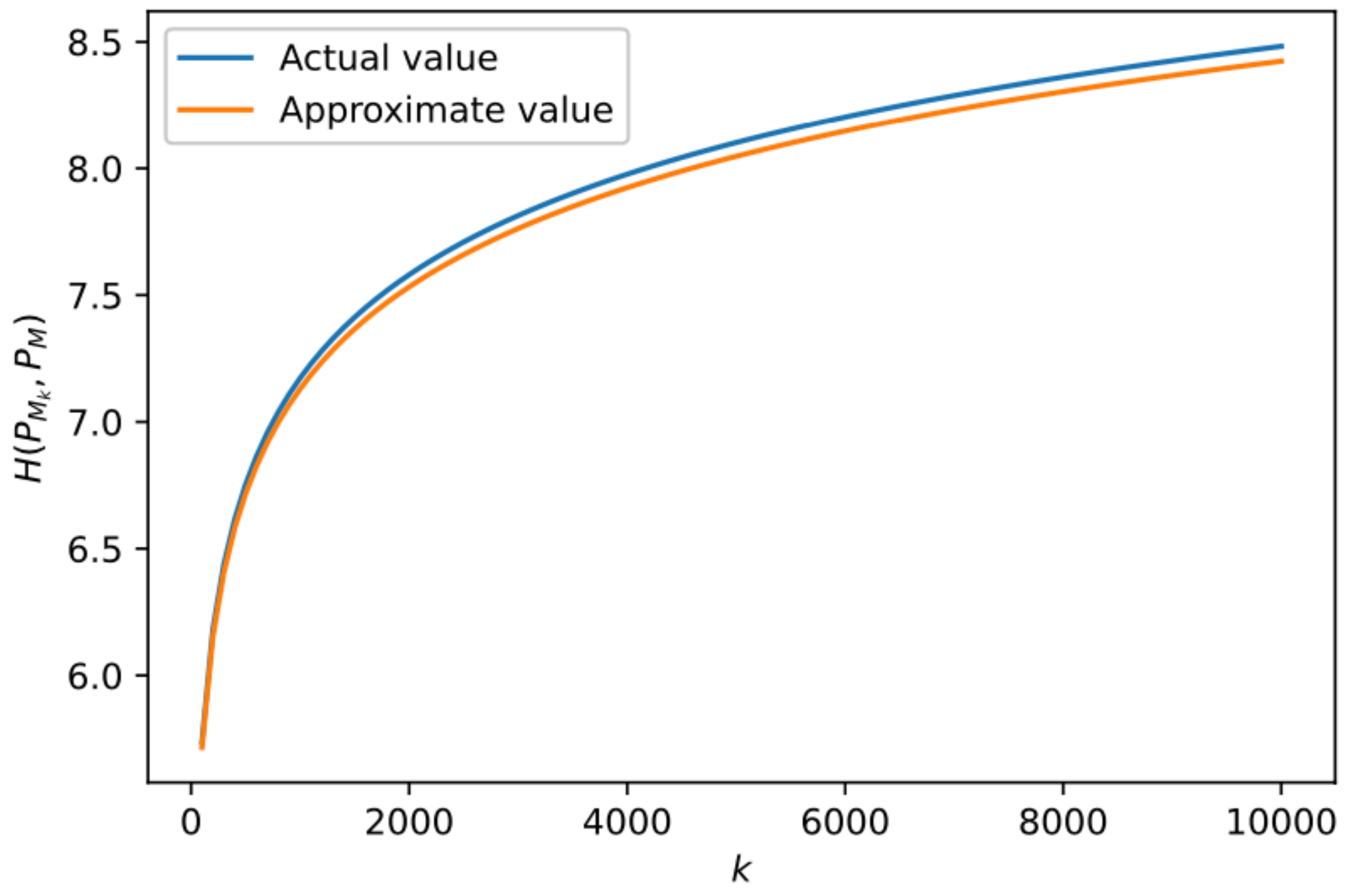
Theoretical Analysis

Theorem: Let P_M be the model distribution satisfying Zipf's law with vocabulary of size N.
and let P_{M_k} be the model distribution obtained using top-k sampling.
Then, for $1 < s \leq \frac{1}{\ln 2}$, $H(P_{M_k}, P_M)$ can be approximated as

$$H(P_{M_k}, P_M) \approx \frac{b_1 \epsilon}{b_3} \left(1 - \frac{b_2 b_3 (\ln k + \frac{1}{\epsilon}) - b_1}{b_1 (b_3 k^{\epsilon} - 1)} \right) + \log H_{N,s},$$ where

$$\epsilon = s - 1 b_1 = s \left(\frac{\log 2}{2^{1+\epsilon}} + \frac{\log 3}{3^{1+\epsilon}} + \frac{1}{\epsilon (\ln 2) 3^{\epsilon}} \left(\ln 3 + \frac{1}{\epsilon} \right) \right), b_2 = \frac{s}{\epsilon \ln 2},$$

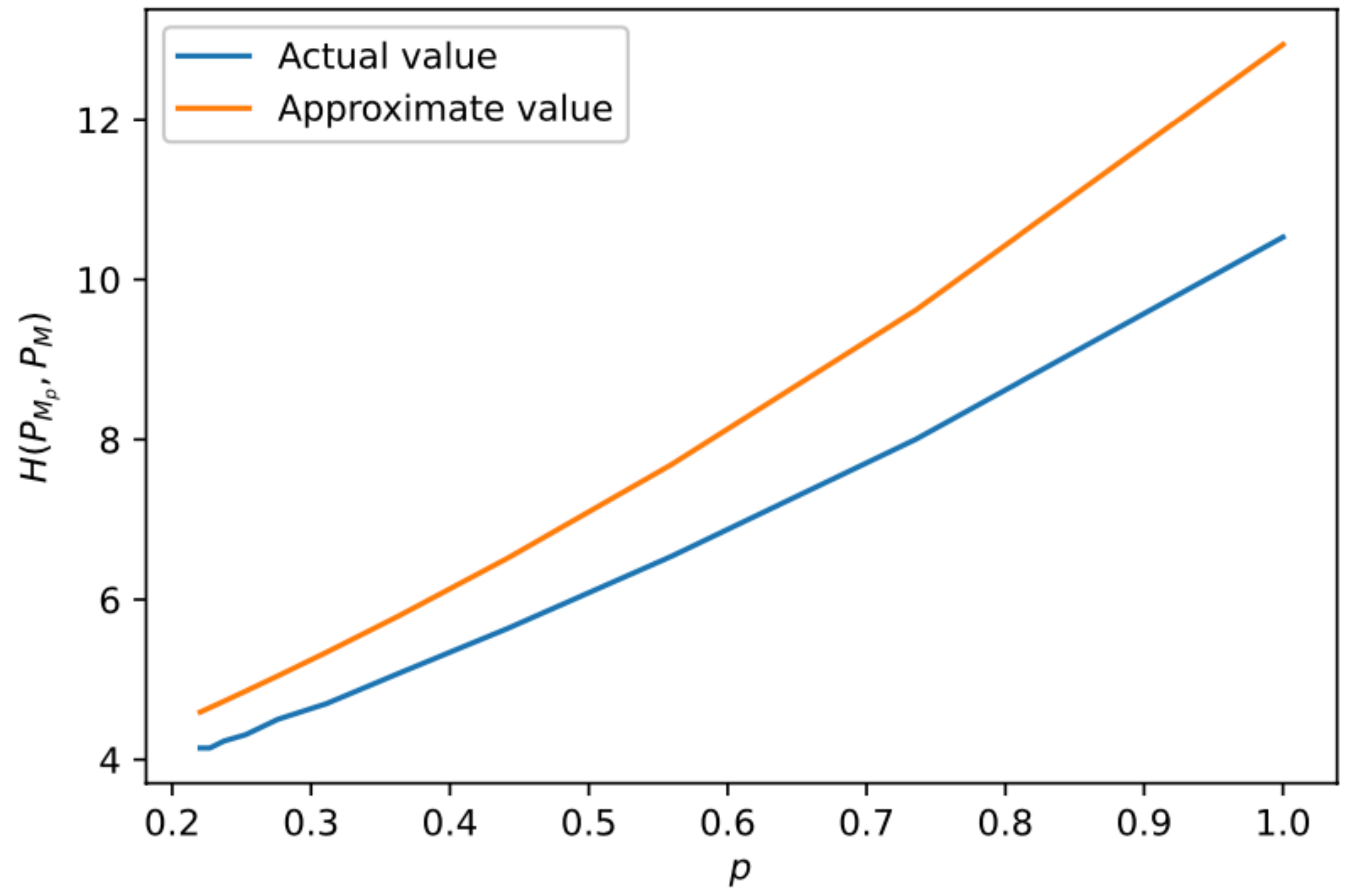
and $b_3 = 1 + 0.7\epsilon$ are constants.



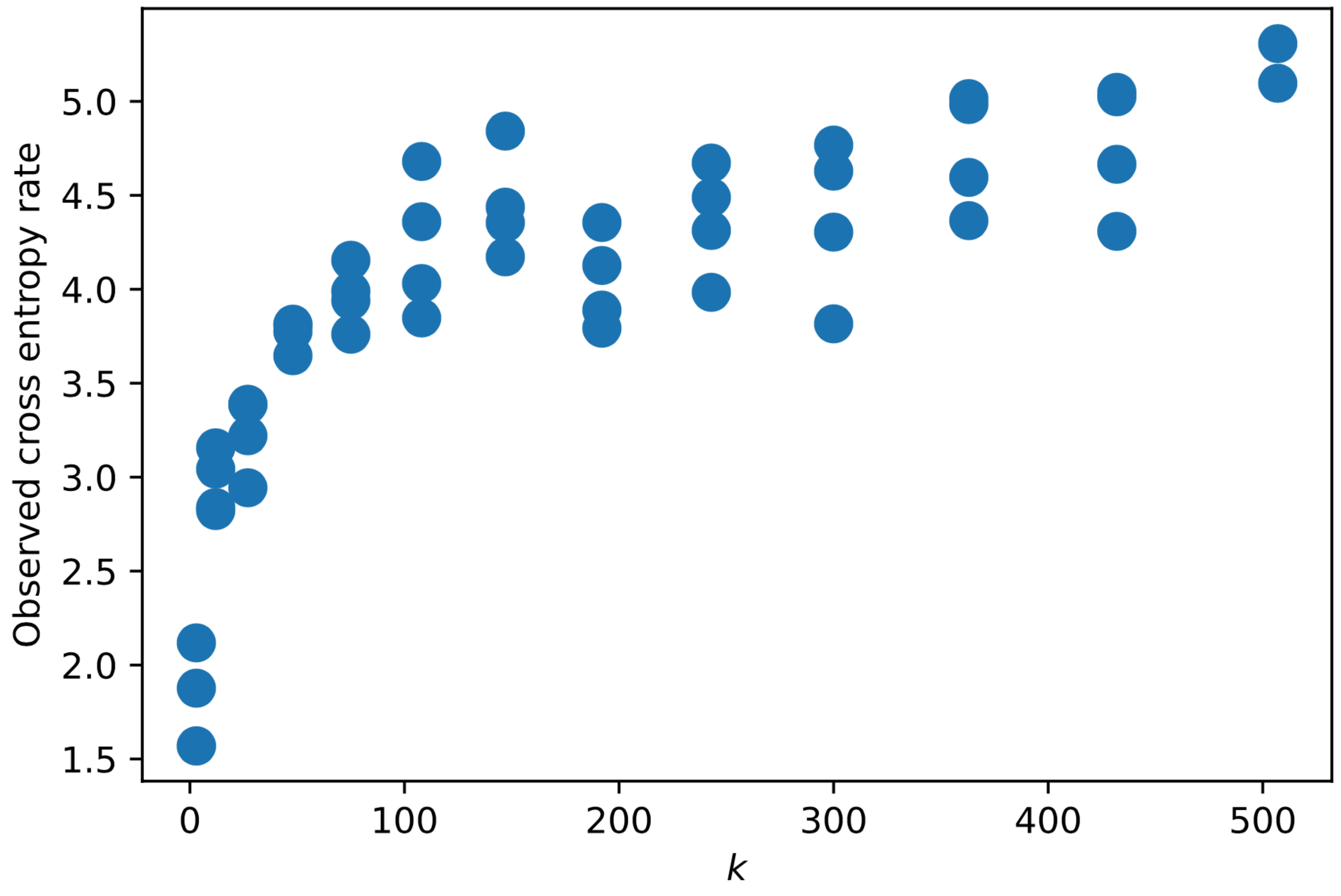
Theoretical Analysis

Theorem: Let P_M be the model distribution satisfying Zipf's law with vocabulary of size N and let $P_{M_{k(p)}}$ be the model distribution obtained using top-p sampling where $k(p)$ is the minimum value of k satisfying $\frac{1}{H_{N,s}} \sum_{i=1}^{k(p)} \frac{1}{i^s} \geq p$. Then, for $1 < s \leq \frac{1}{\ln 2}$, $H(P_{M_p}, P_M)$ can be approximated as

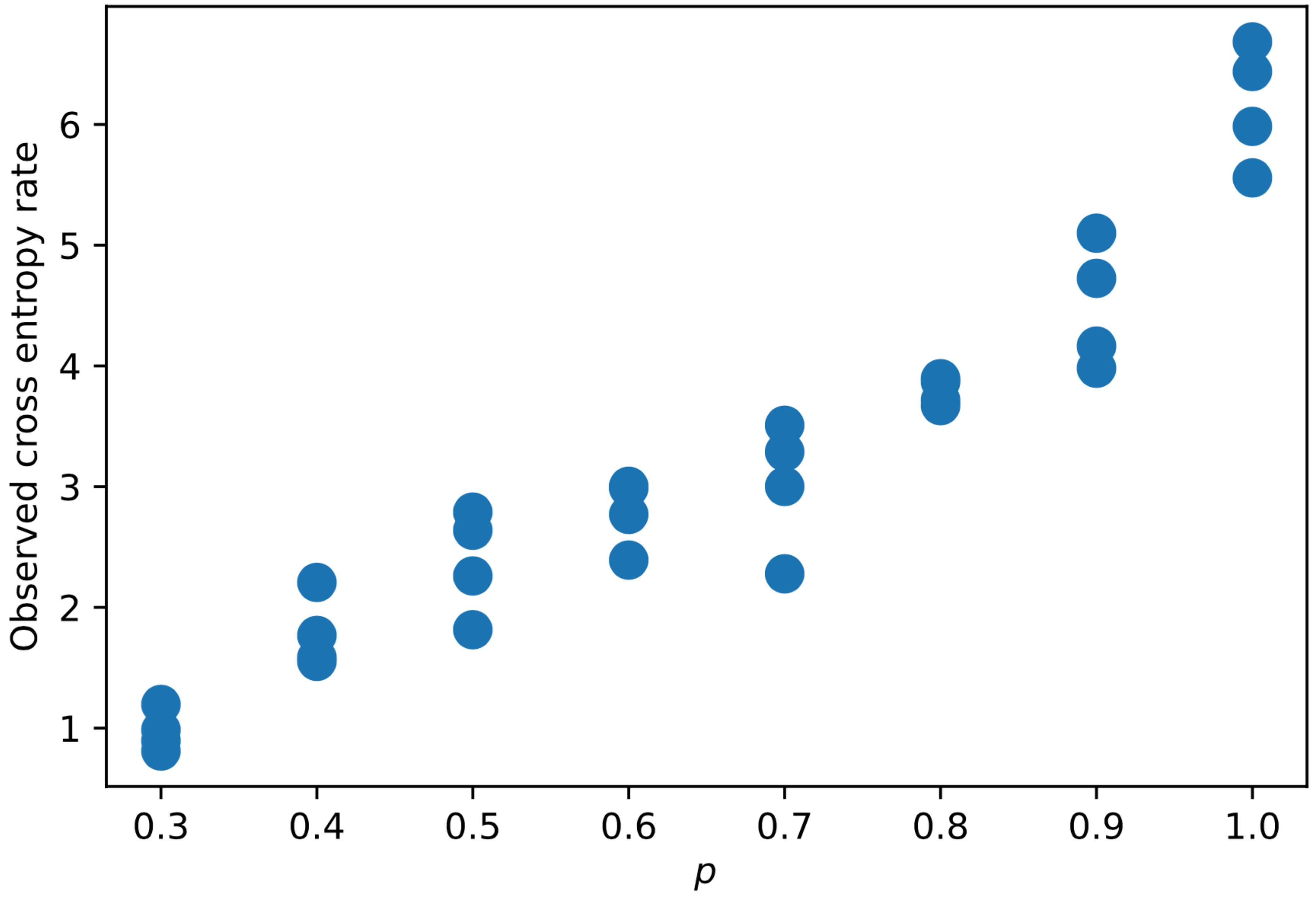
$$H(P_{M_p}, P_M) \approx \frac{s}{2 \ln 2} \left(pH_{N,s} + \epsilon p^2 H_{N,s}^2 \right) + \log H_{N,s} .$$



Experimental Analysis



(a) Top-k sampling



(b) Top-p sampling

Mirostat Sampling

Mirostat Sampling Algorithm

Algorithm 1: Adaptive top-k sampling for perplexity control

Target cross entropy τ, maximum cross entropy $\mu = 2 * \tau$, learning rate η

while *more words are to be generated* **do**

Compute $\hat{s}$ from (40): $\frac{\sum_{i=1}^{N-1} t_i b_i}{\sum_{i=1}^{N-1} t_i^2}$

Compute k from (41): $k = \left(\frac{\hat{\epsilon} 2^{\mu}}{1 - N^{-\hat{\epsilon}}} \right)^{\frac{1}{\hat{s}}}$

Sample the next word X using top-k sampling

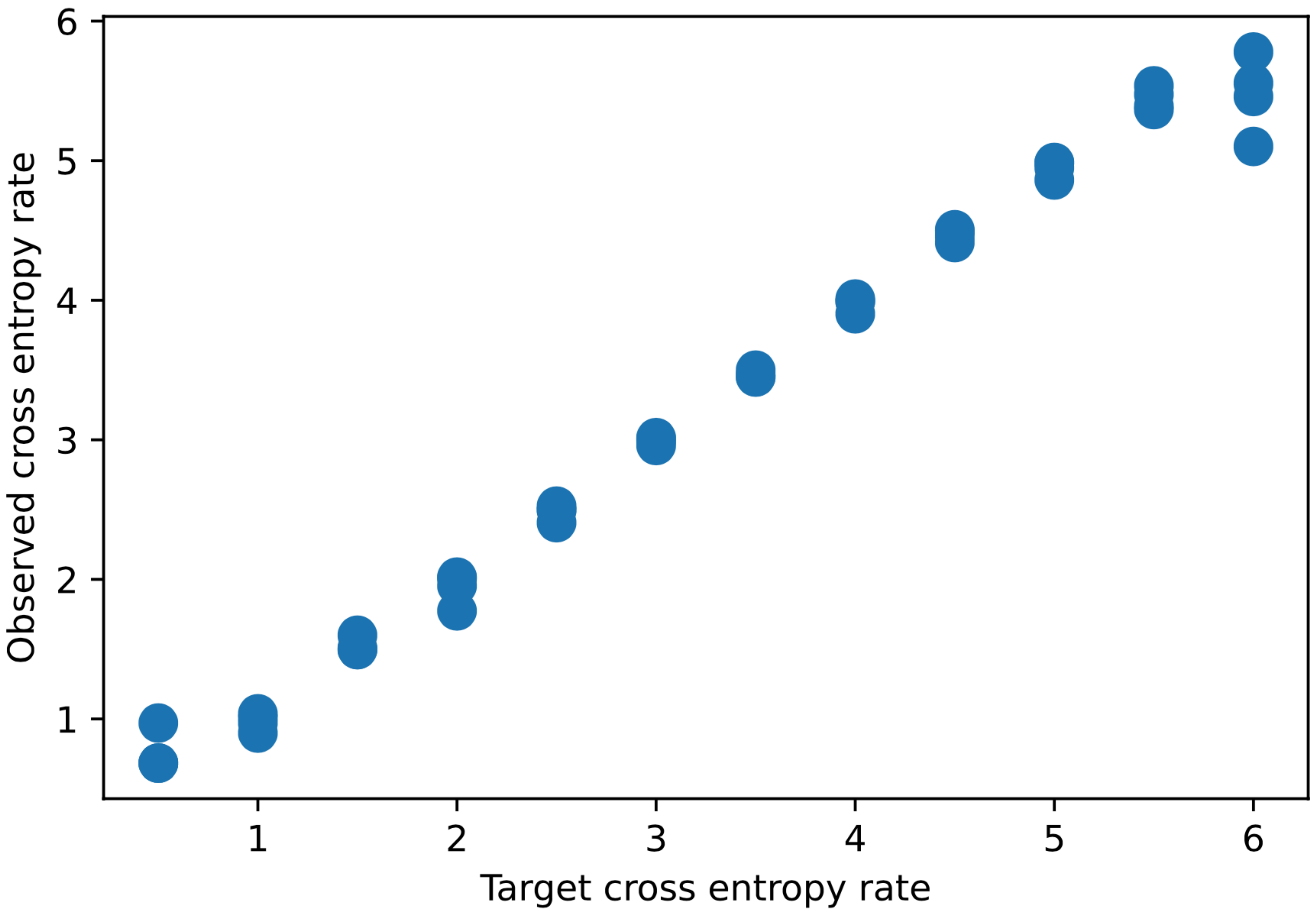
Compute error: $e = \mathfrak{S}(X) - \tau$

Update μ: $\mu = \mu - \eta * e$

end

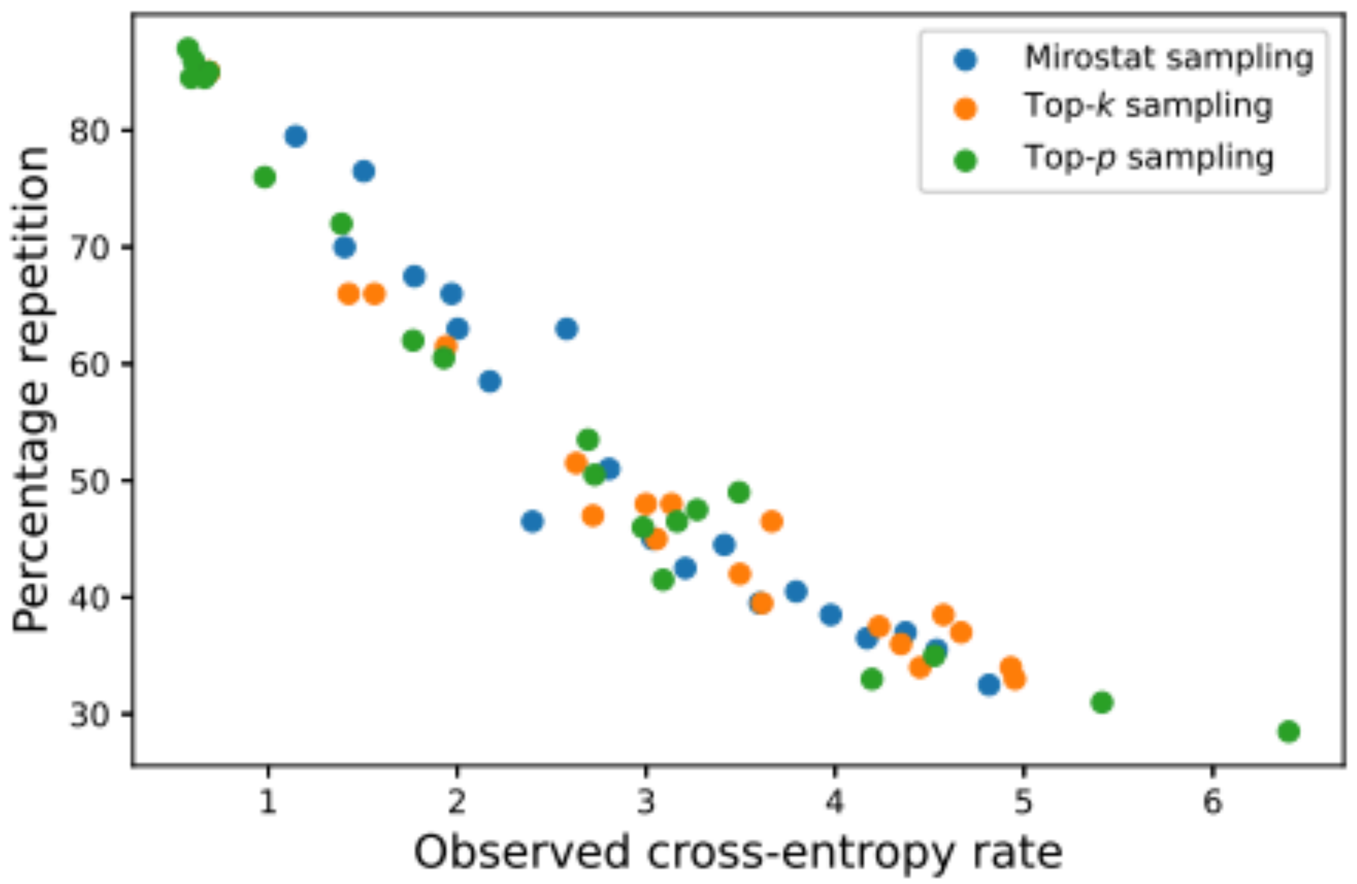
Experiments

Controlled Cross-entropy

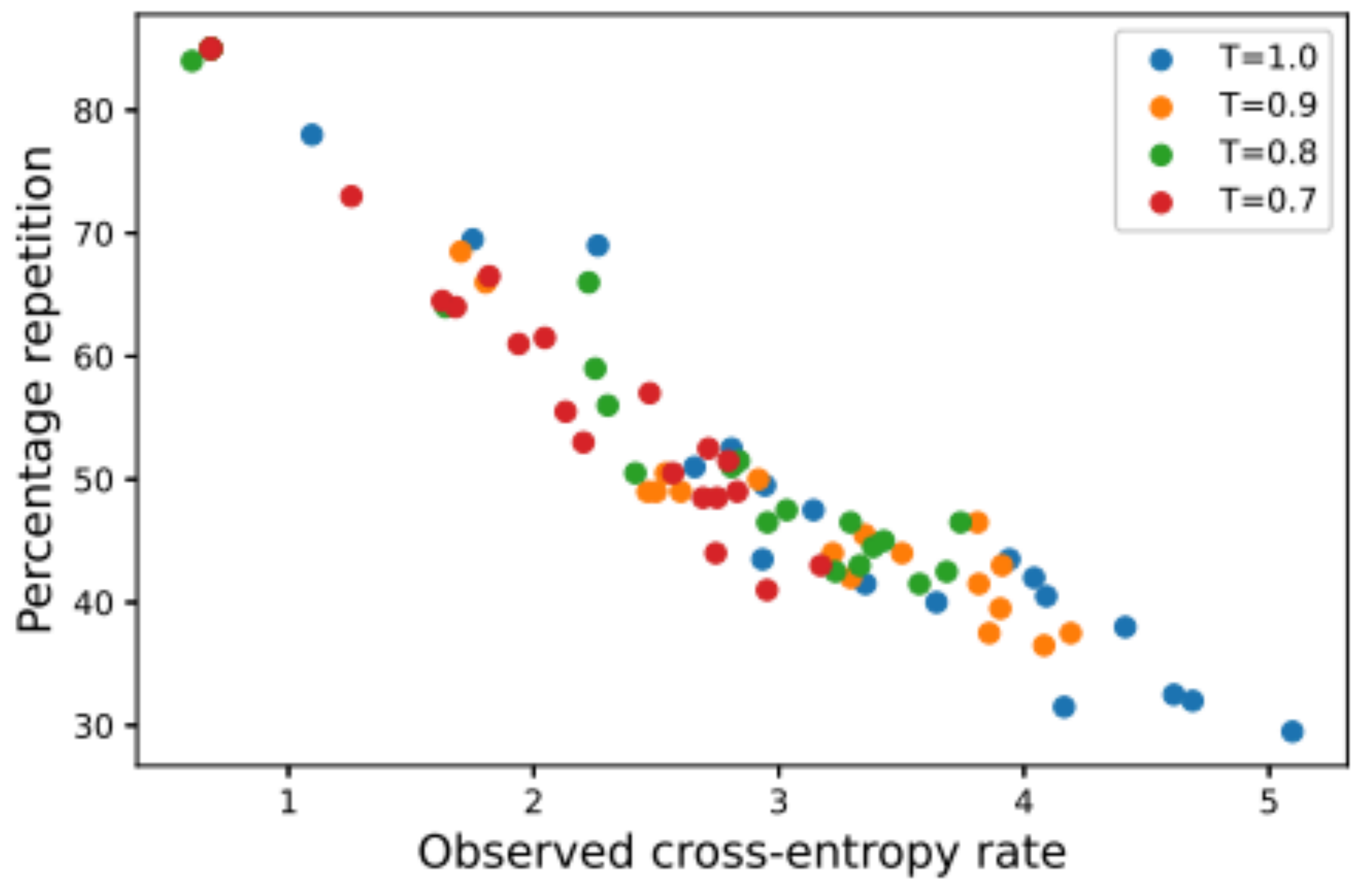


(c) Mirostat sampling

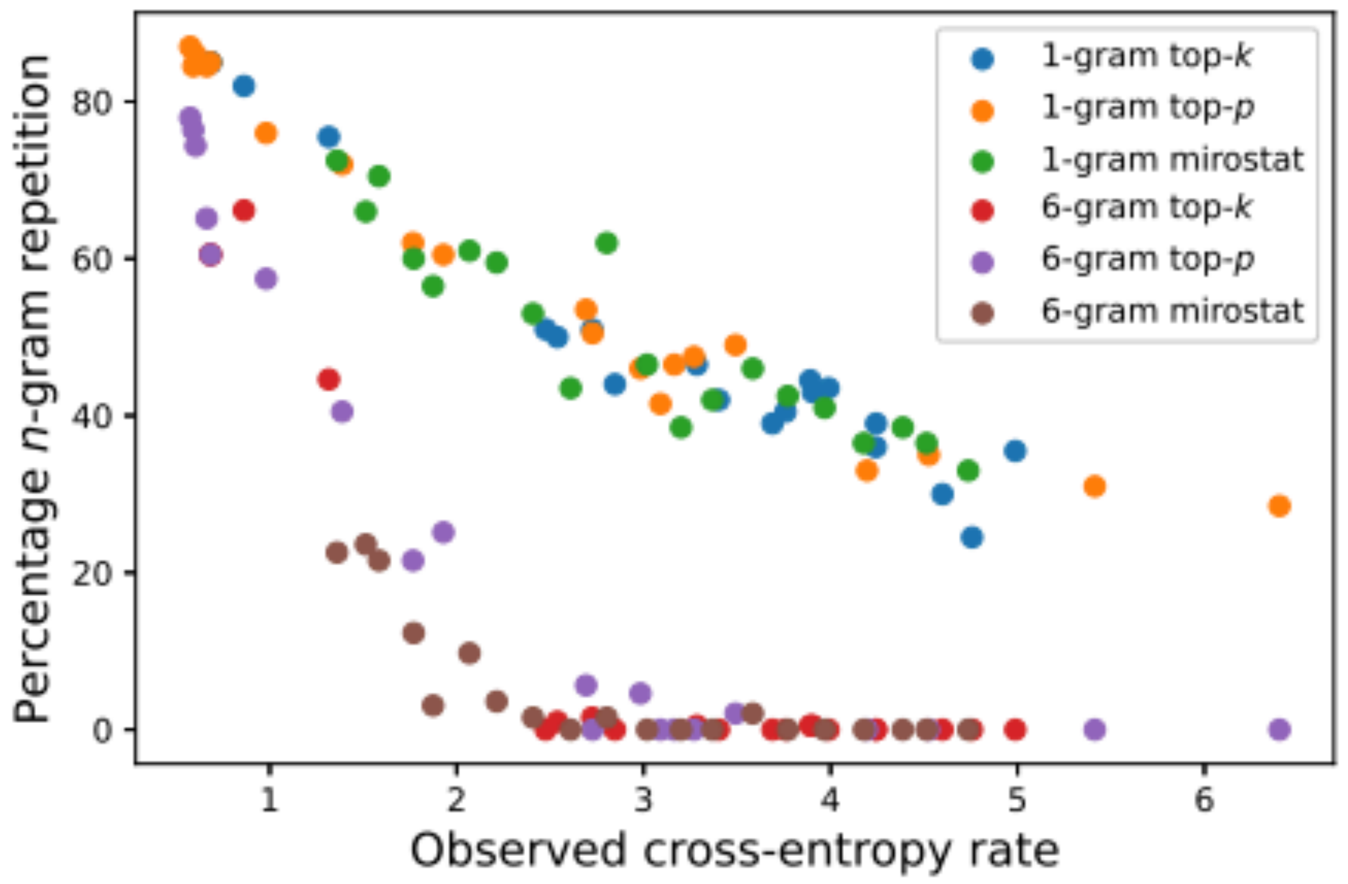
Repetition Control



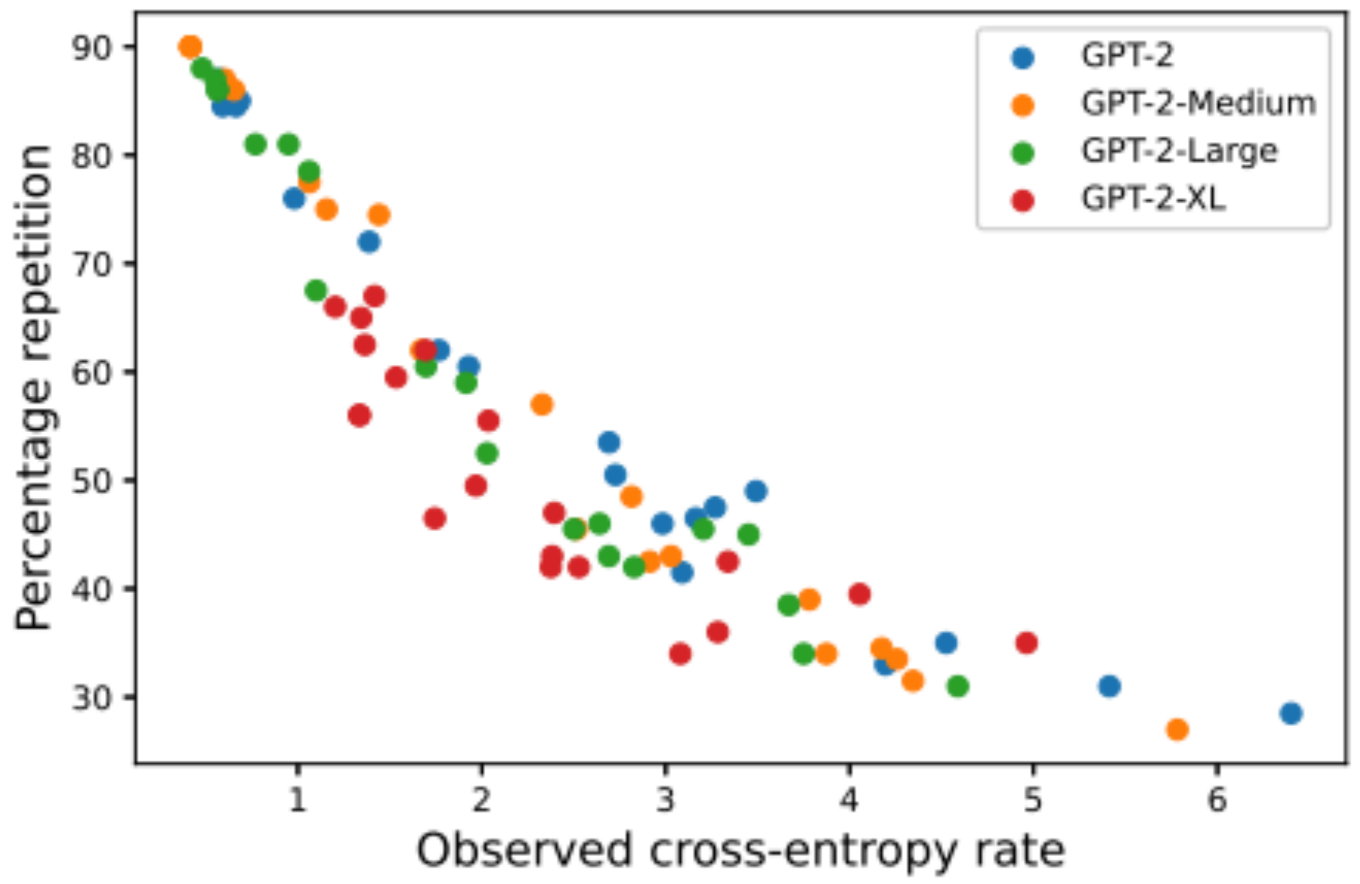
(a)



(b)

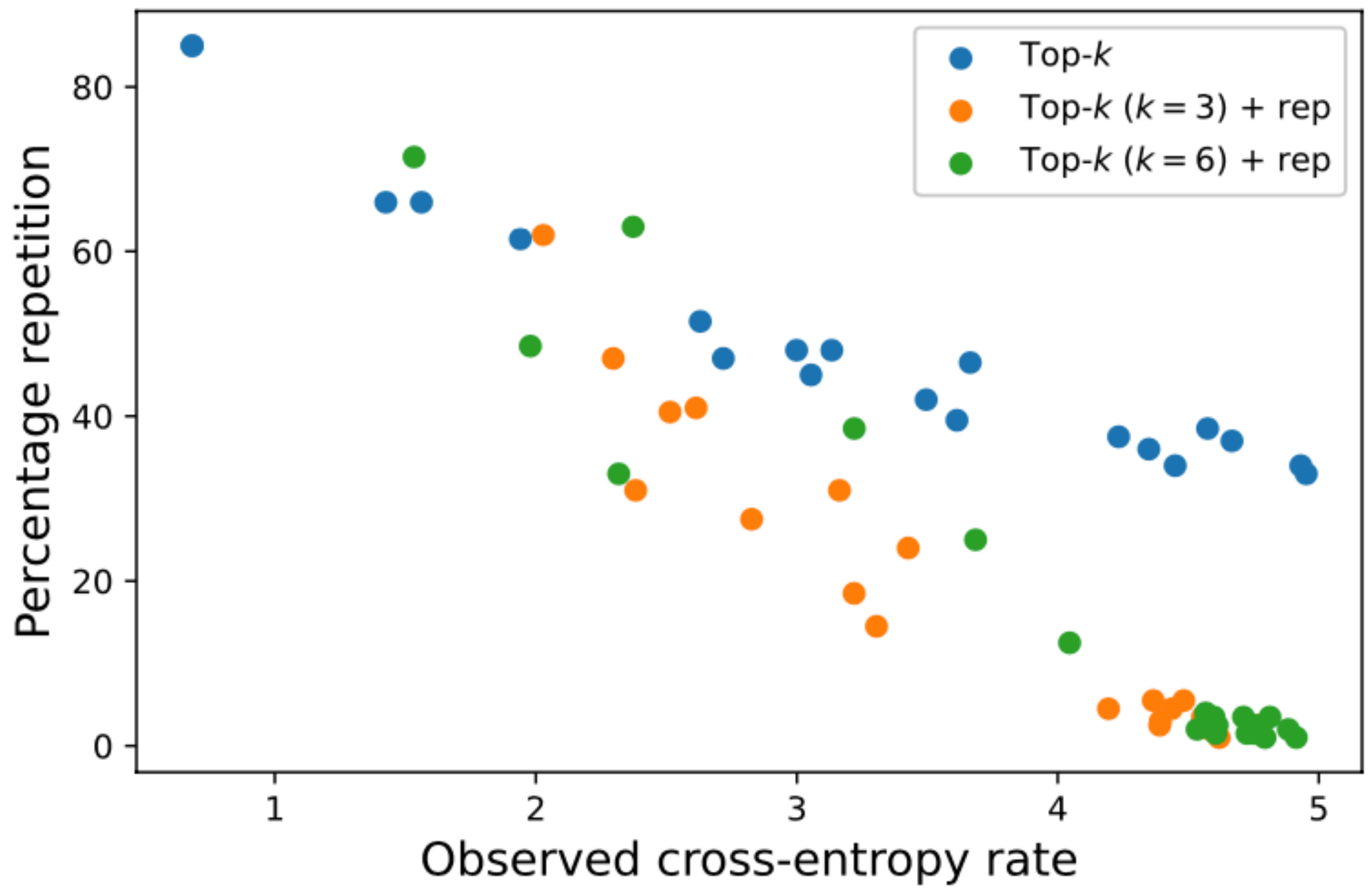


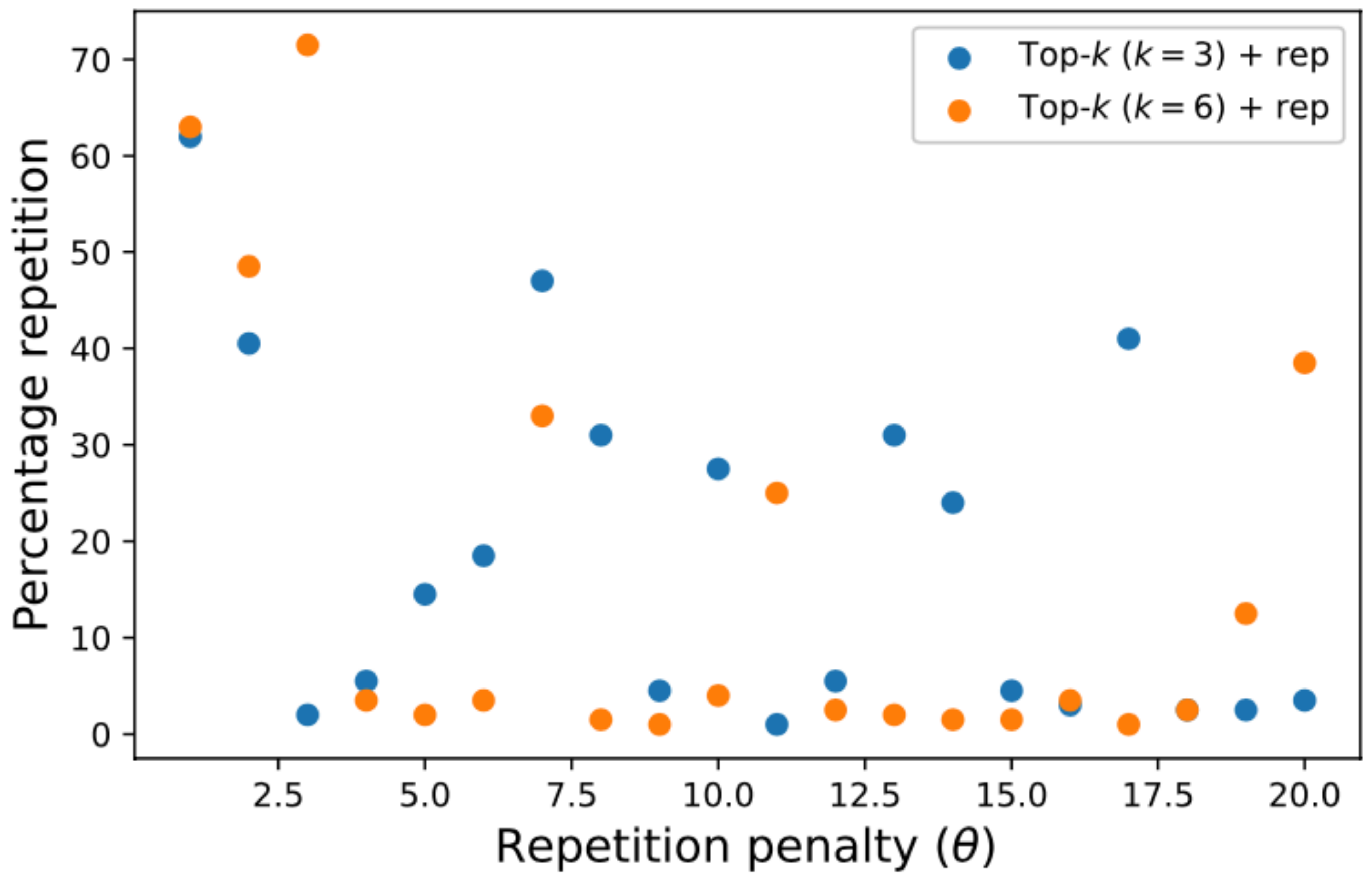
(c)



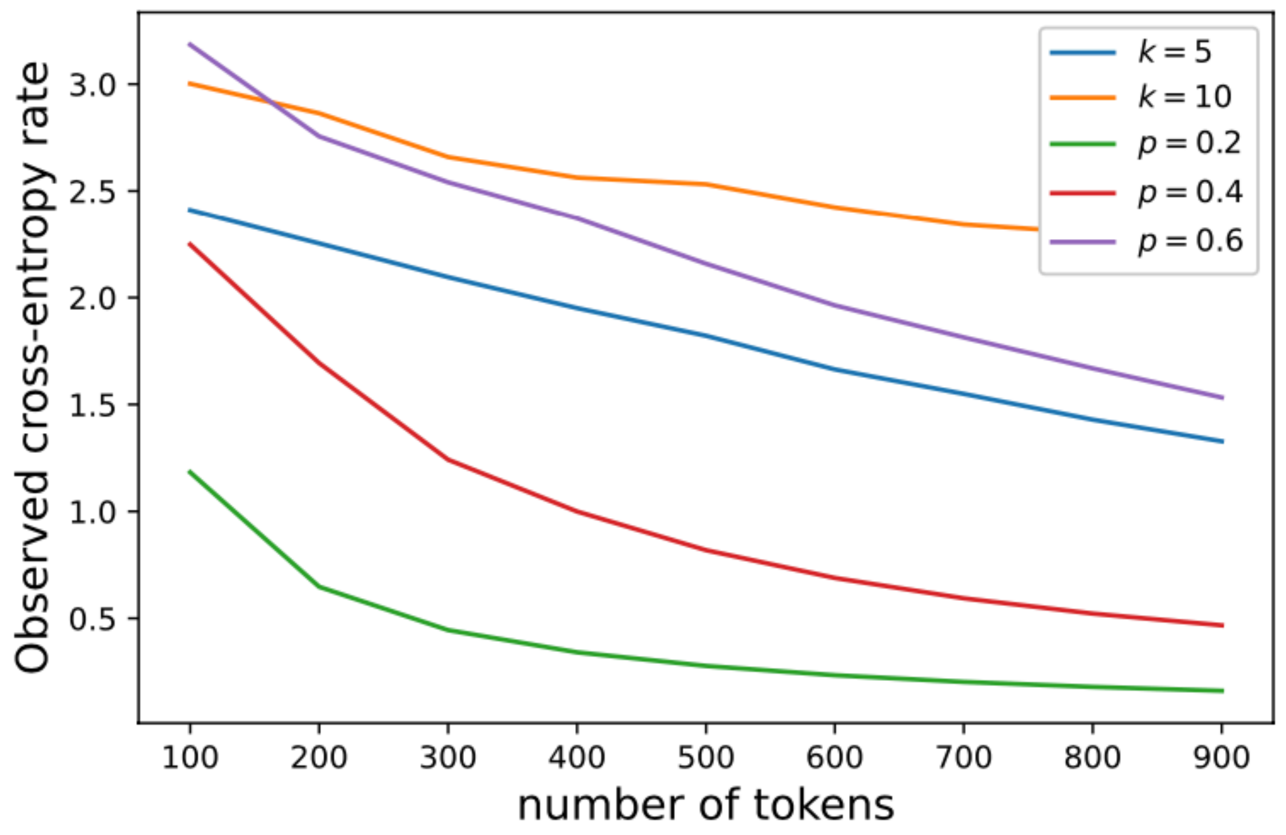
(d)

Comparison to Repetition Penalty

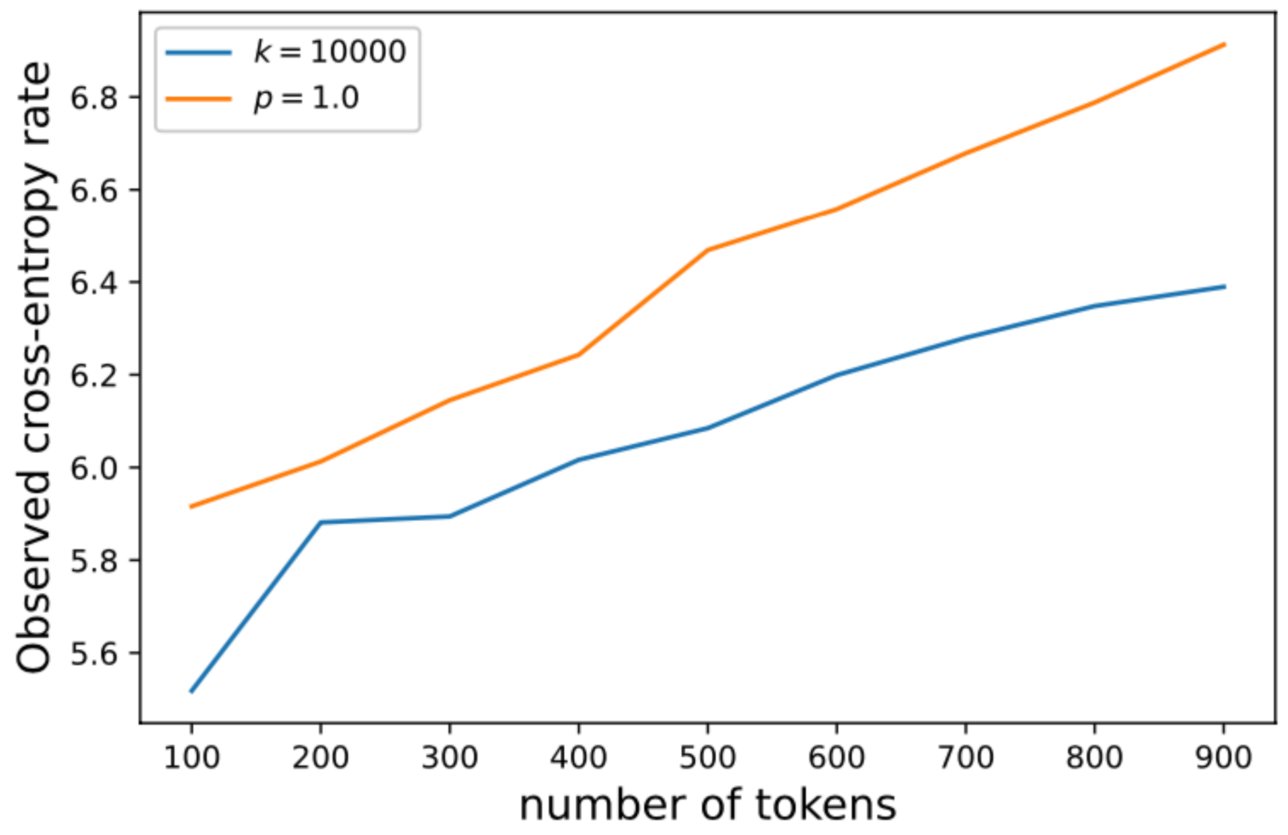




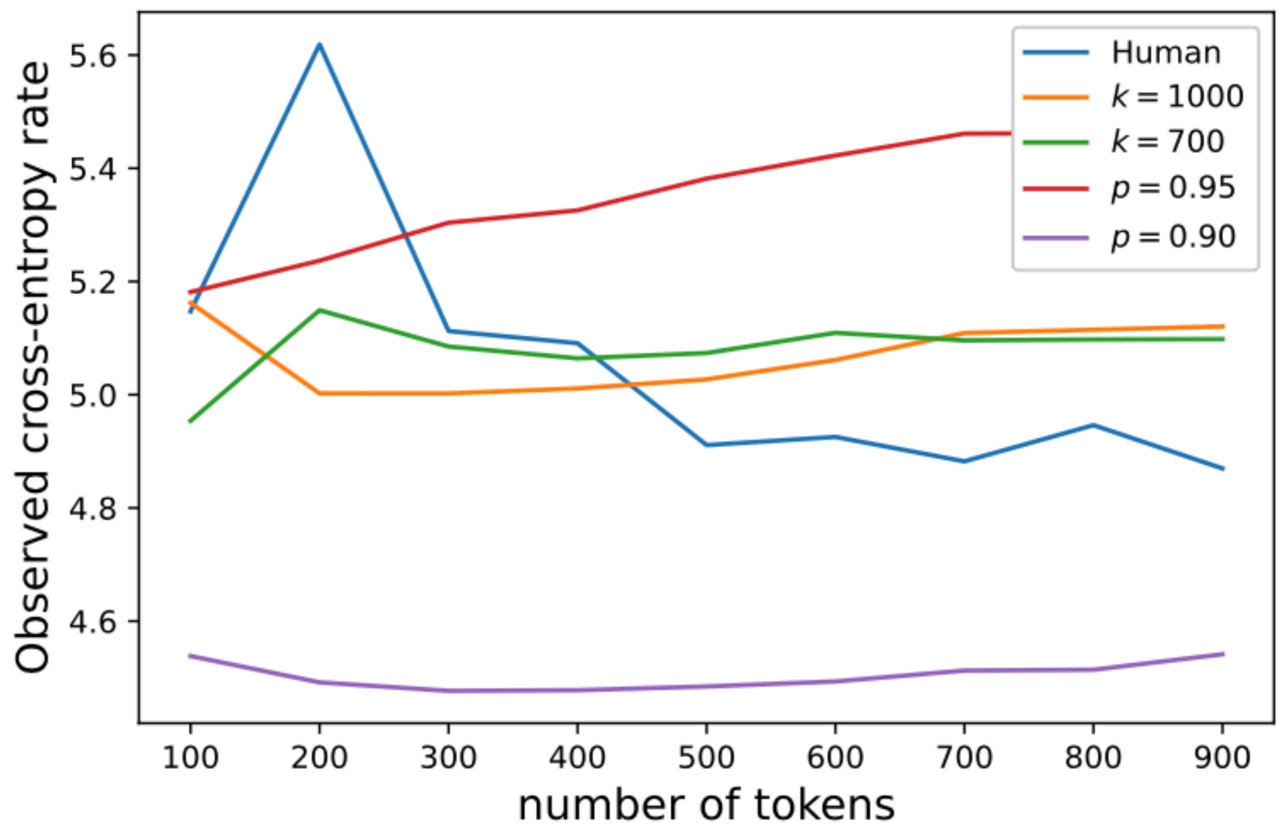
Cross-entropy Control over Length



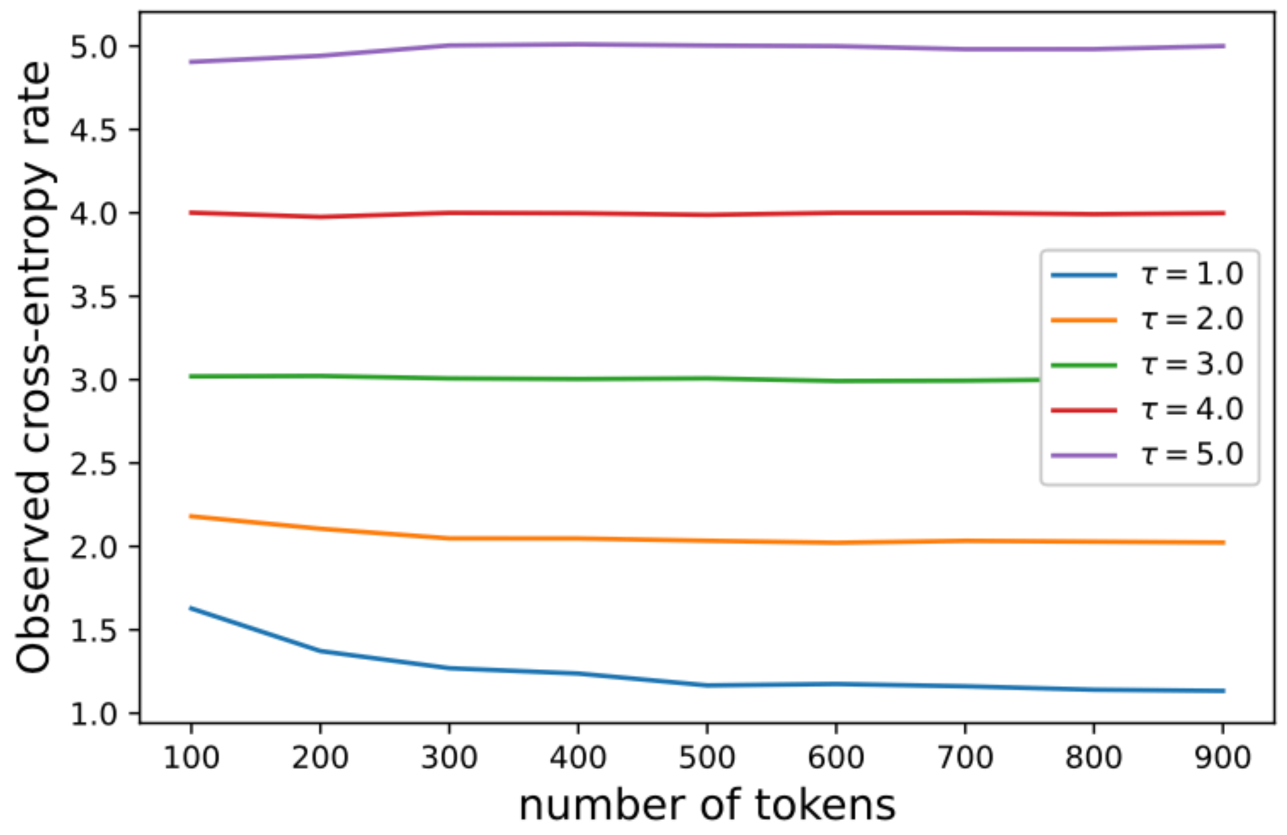
(a)



(b)

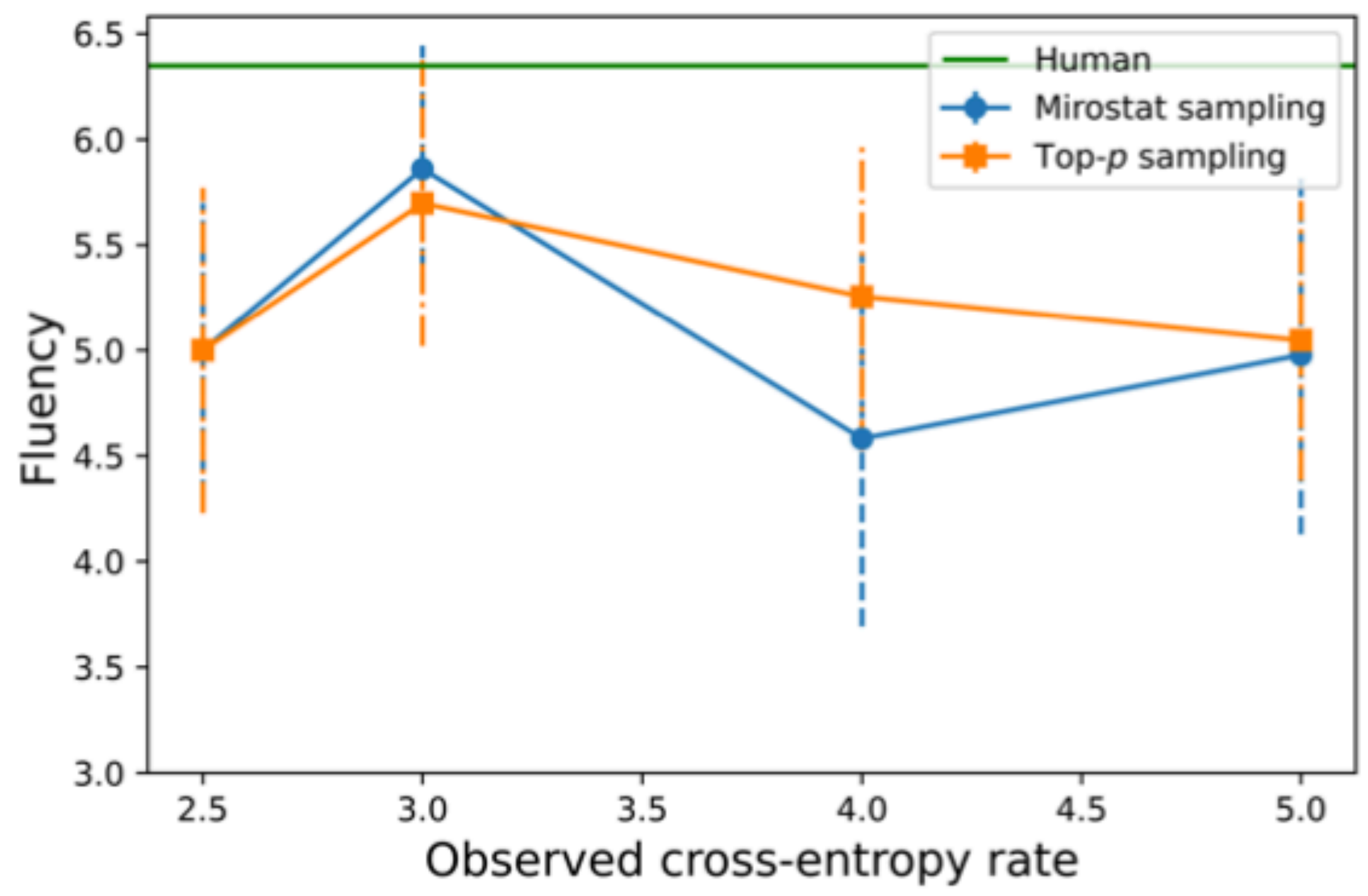


(c)

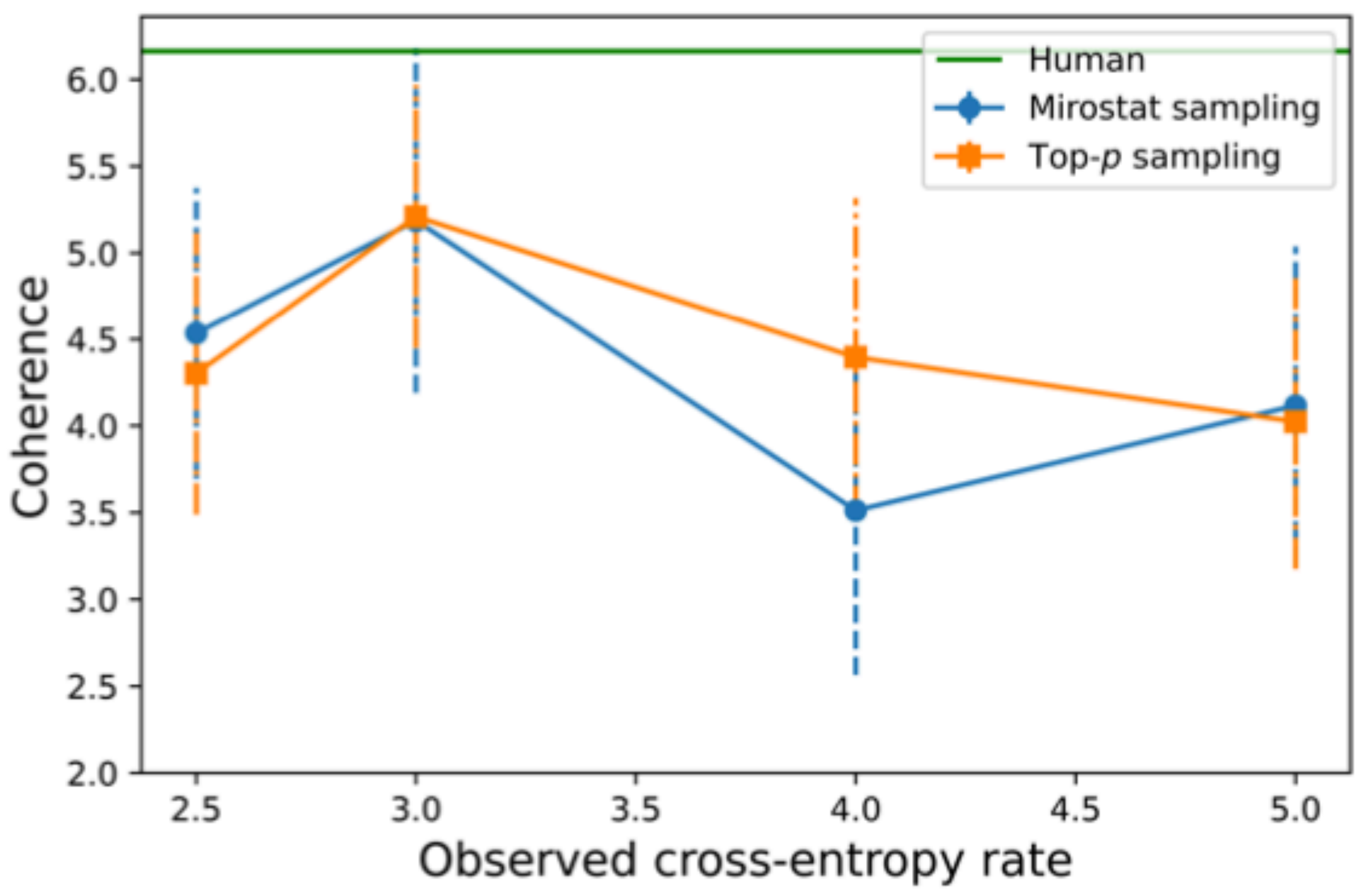


(d)

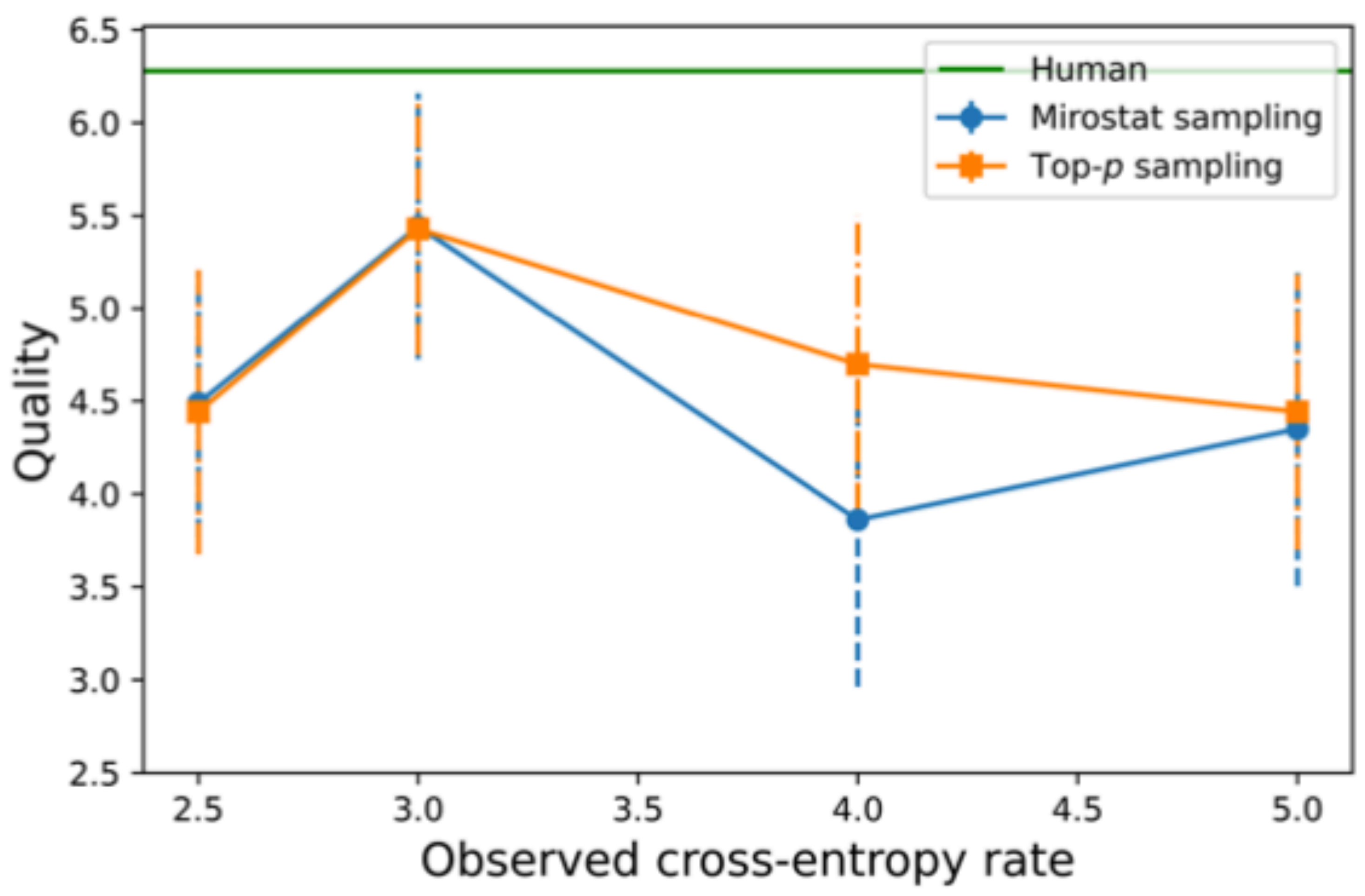
Human Evaluations



(a)

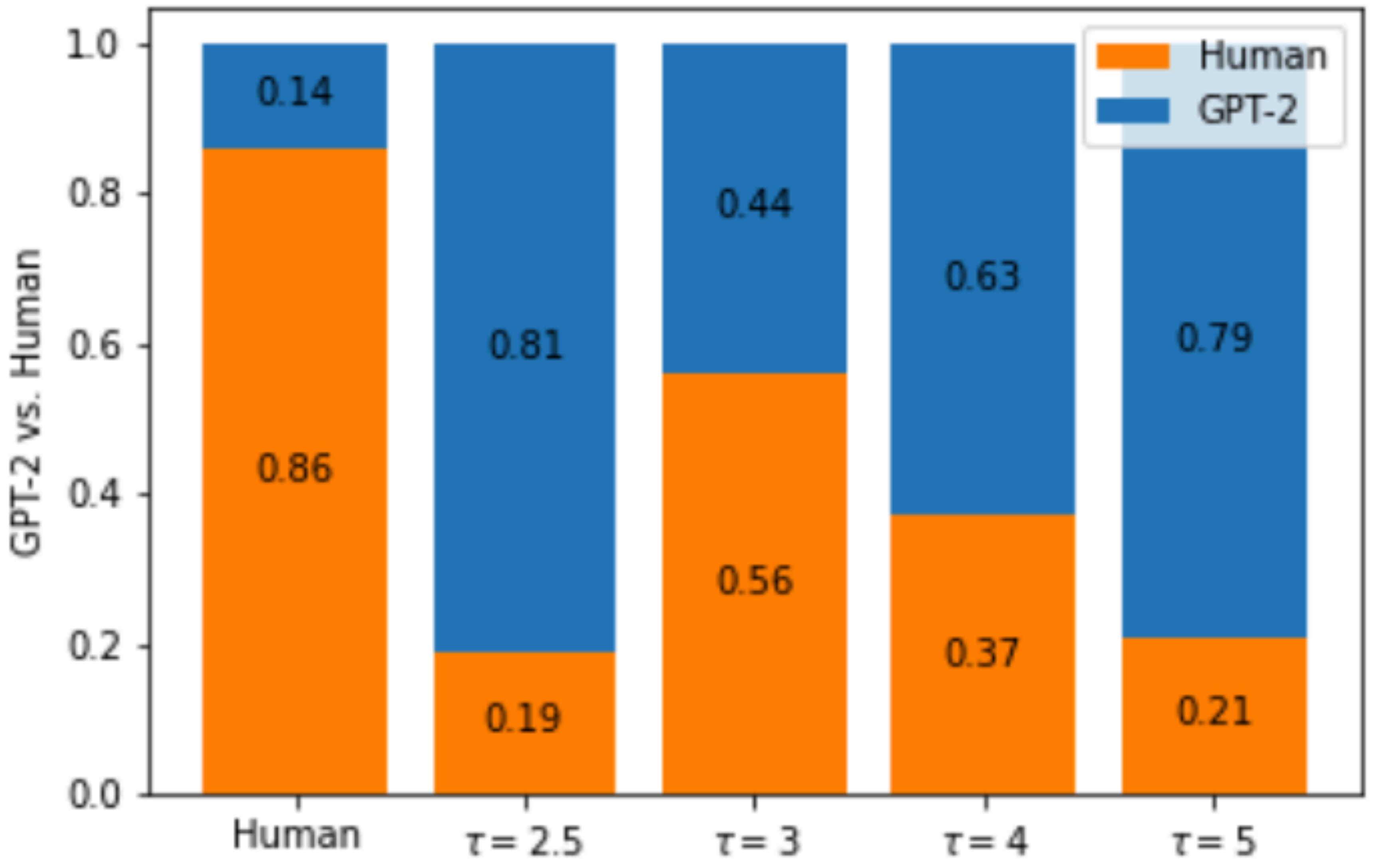


(b)

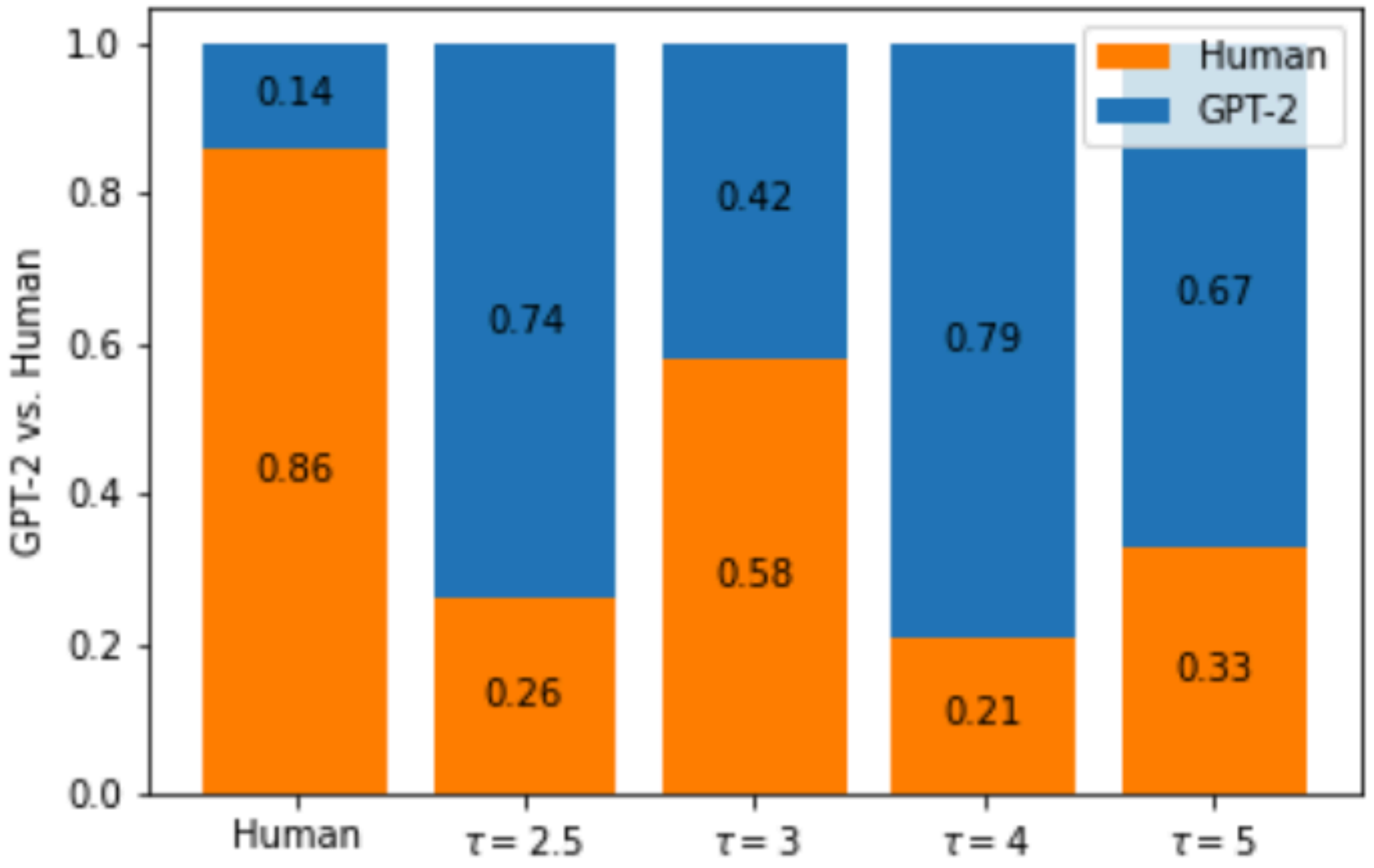


(c)

Human Evaluations



(d)



(e)

Thank you!