# The Curious Case of Neural Text Degeneration

## Decoding Strategies for Large Language Models

Elhadi Chiter, Solomon Harvey, Marc Kaspar, Aziz Agrebi

Université Paris-Dauphine - PSL

December 2024

# Introduction

**Research Paper:**
"The Curious Case of Neural Text Degeneration" (Holtzman et al., 2020)

**Key Points:**

- Focus on decoding strategies for Large Language Models (LLMs).
- Introduces *Nucleus Sampling* (top-$p$ sampling) as a new approach.
- Addresses issues of text degeneration: blandness, incoherence, and repetition.

# Decoding Strategies Overview

**Existing Methods:**

- **Maximization-Based Methods:**
  - Greedy Search, Beam Search
  - Issues: Lack of diversity, high repetition
- **Sampling-Based Methods:**
  - Pure Sampling, Sampling with Temperature, Top-$k$ Sampling
  - Issues: Difficulty balancing diversity and coherence

# Nucleus Sampling (Top-$p$ Sampling)

**Key Idea:**

- Dynamically selects tokens comprising the top-$p$% of the probability mass.
- Adjusts the sampling set based on probability distribution.

**Mathematical Definition:**

$$\sum_{x \in V^{(p)}} P(x|x_{1:i-1}) \geq p$$

**Advantages:**

- Dynamic adaptation to probability distribution.
- Balances coherence and diversity better than Top-$k$ and Temperature Sampling.

# Comparison of Sampling Methods

**Top-$k$ Sampling:**

- Fixed number of tokens considered.
- Sensitive to distribution shape.
- May ignore key tokens or include unrelated ones.

**Nucleus Sampling:**

- Considers tokens based on cumulative probability.
- Adapts dynamically to model confidence.
- Avoids issues with fixed $k$ selection.

# Comparison of Sampling Methods

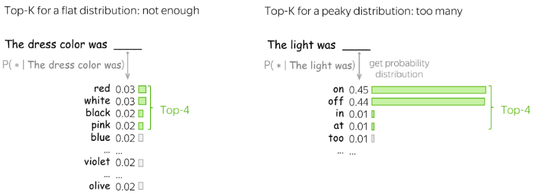


Figure: Top-$k$ Sampling

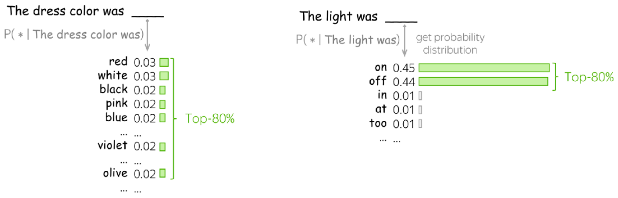


Figure: Nucleus Sampling

# Limitations and Open Questions

**Limitations of Nucleus Sampling:**

- Threshold $p$ selection is non-trivial.
- Larger $p$ values may lead to larger sampling sets.

**Open Questions:**

- How to efficiently choose $p$?
- Can a metric guide the selection of $p$ for specific applications?

# Metrics Overview

**Metrics Defined:**

- **Perplexity:** Measures model confidence in predicting text.
- **Self-BLEU:** Assesses diversity across generated outputs.
- **Repetition:** Identifies repetitive patterns within outputs.
- **Zipf Coefficient:** Evaluates adherence to Zipf's law of token frequency.

# Perplexity

**Definition:**

$$\text{Perplexity}(T) = \exp\left(\frac{1}{N}\sum_{i=1}^{N} -\log P(w_i \mid w_1, \ldots, w_{i-1})\right)$$

- $T$: Pre-written text of length $N$.
- Measures how well the model predicts $T$.
- Low perplexity indicates high confidence; high perplexity indicates confusion.

**Experiment:**

- Computed on 10,000 tokens from WikiText dataset.

# Self-BLEU

**Definition:**

$$\text{Self-BLEU}(G_1, \ldots, G_m) = \frac{1}{m} \sum_{i=1}^{m} \text{BLEU}\left( G_i, \bigcup_{j \neq i} G_j \right)$$

- $G_1, \ldots, G_m$: Generated outputs for a prompt.
- Measures diversity of outputs.
- Score close to 0: High diversity; score close to 1: Low diversity.

**Experiment:**

- Used standard Self-BLEU instead of Self-BLEU4 for precision.
- Averaged over all prompts for each decoding strategy.

# Repetition

**Definition:**

$$\text{Repetition}(G, W) = 100 \times \frac{\text{RepeatedTokens}(G, W)}{|G|}$$

- $G$: Output text; $W$: Window size.
- Identifies repeated patterns of $n$-grams within the last $W$ tokens.

**Experiment:**

- Computed repetition for each output.
- Averaged over outputs for each decoding strategy.

# Zipf Coefficient

**Definition:**

$$f(w) \simeq \frac{1}{r(w)^{s(G)}}$$

- $f(w)$: Frequency of token $w$.
- $r(w)$: Rank of token based on frequency.
- $s(G)$: Zipf coefficient of the output $G$.

**Experiment:**

- Evaluates adherence to natural linguistic patterns.
- Deviation indicates unnatural token distribution.

# Implementation of Code

**Introduction:**

- Implemented metrics for different decoding strategies.
- Used Mistral 7B model for generating outputs.
- Code available on GitHub: github.com/spharvey99/llm-project.

**Procedure:**

- Generated outputs for 5 prompts using each decoding strategy.
- Computed metrics for each strategy using outputs of up to 200 words.

# Results of the experiment

| Method | Perplexity | Self-BLEU4 | Zipf Coef | Repetition % | HUSE |
|---|---|---|---|---|---|
| Human | 12.38 | 0.31 | 0.93 | 0.28 | - |
| Greedy | 1.50 | 0.50 | 1.00 | 73.66 | - |
| Beam, $b = 16$ | 1.48 | 0.44 | 0.94 | 28.94 | - |
| Stoch. Beam, $b = 16$ | 19.20 | 0.28 | 0.91 | 0.32 | - |
| Pure Sampling | 22.73 | 0.28 | **0.93** | 0.22 | 0.67 |
| Sampling, $t = 0.9$ | 10.25 | 0.35 | 0.96 | 0.66 | 0.79 |
| Top-$k = 40$ | 6.88 | 0.39 | 0.96 | 0.78 | 0.19 |
| Top-$k = 640$ | 13.82 | **0.32** | 0.96 | **0.28** | 0.94 |
| Top-$k = 40$, $t = 0.7$ | 3.48 | 0.44 | 1.00 | 8.86 | 0.08 |
| Nucleus, $p = 0.95$ | **13.13** | **0.32** | 0.95 | 0.36 | **0.97** |

Figure: Results of the paper

| Method | Perplexity | Self-BLEU | Zipf | Repetition (%) |
|---|---|---|---|---|
| Beam Search ($b = 4$) | 1.6987 | 1.0000 | 0.6274 | 0.0 |
| Pure Sampling | 19.0512 | 0.3907 | 0.9298 | 0.0 |
| Temperature ($t = 0.9$) | 19.9394 | 0.4228 | 0.9486 | 0.0 |
| Top-k ($k = 640$) | 12.0812 | 0.4380 | 0.9475 | 0.0 |
| Top-k with Temp. ($k = 40, t = 0.7$) | 6.0772 | 0.5021 | 0.9958 | 0.0 |
| Nucleus Sampling ($p = 0.95$) | 8.4538 | 0.4690 | 0.9672 | 0.0 |

Figure: Results of the experiment

# Conclusion

**Summary:**

- Neural text degeneration remains a critical challenge.
- Nucleus Sampling offers a more adaptive approach than previous methods.
- Future work needed to optimize threshold selection and evaluate practical applications.