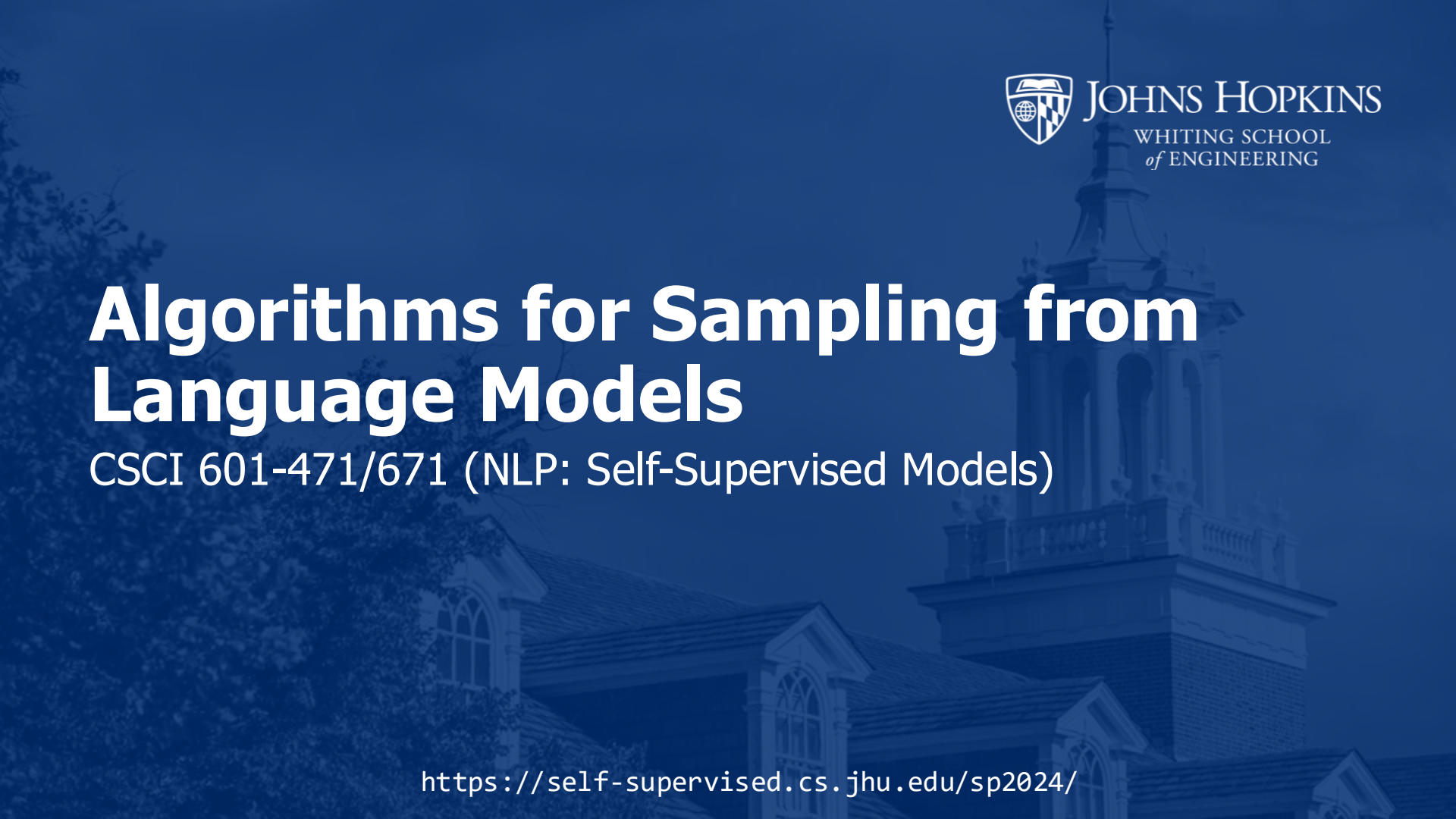



# Algorithms for Sampling from Language Models

CSCI 601-471/671 (NLP: Self-Supervised Models)

https://self-supervised.cs.jhu.edu/sp2024/

# The Sampling Question

How do we generate language from LMs?

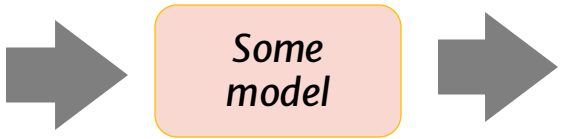
Given:

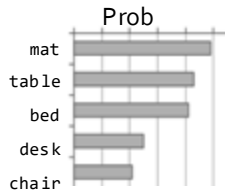
$$\mathbf{P}(X|\, \boldsymbol{x}_{<t}) = \mathbf{P}(X|\, x_1, \ldots, x_{t-1})$$

next word

context

"The cat sat on the [MASK]"



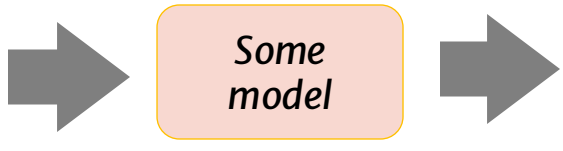



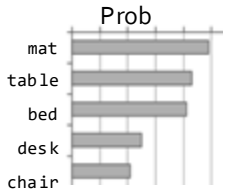
# Greedy Sampling

- **Greedy:** Deterministically choose most probable token according **P**(X| $x_{<t}$)
- Challenge:
  - May repeat itself. "I went to the place that the place that the place that the place ..."
  - Generates boring results — not creative.

$$x = \text{argmax}_{x \in V} \ \mathbf{P}(X = x | \boldsymbol{x}_{<t})$$

"The cat sat on the [MASK]"



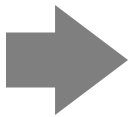

[The Curious Case of Neural Text Degeneration, Holtzman et al., 2020]

# Sampling from the whole distribution

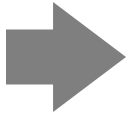
- Sample according to multinomial distribution given by **P**(X| $x_{<t}$)

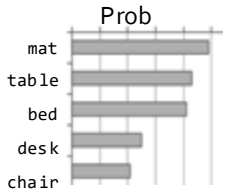
$$x \sim \mathbf{P}(X = x|\boldsymbol{x}_{<t})$$

"The cat sat on the [MASK]"



*Some model*



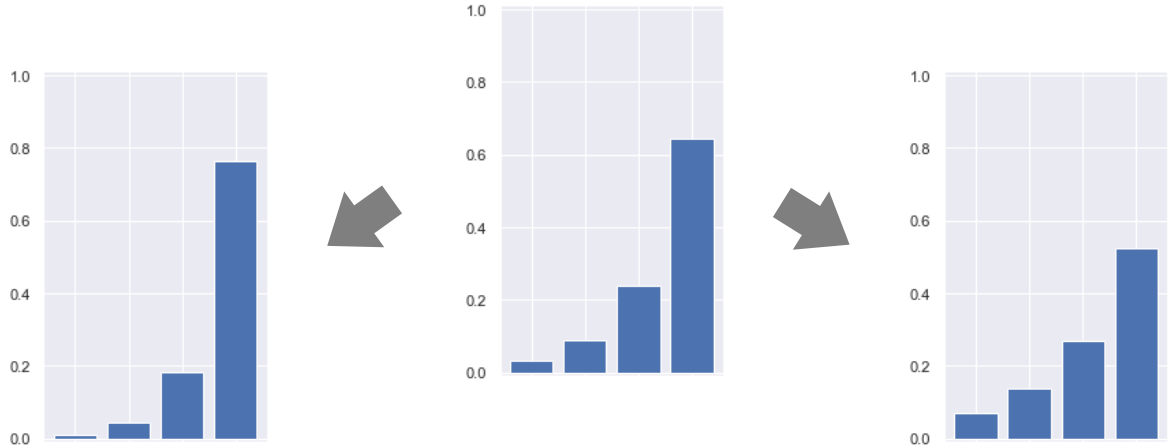



[The Curious Case of Neural Text Degeneration, Holtzman et al., 2020]

# Quiz: Softmax Temperature Parameter

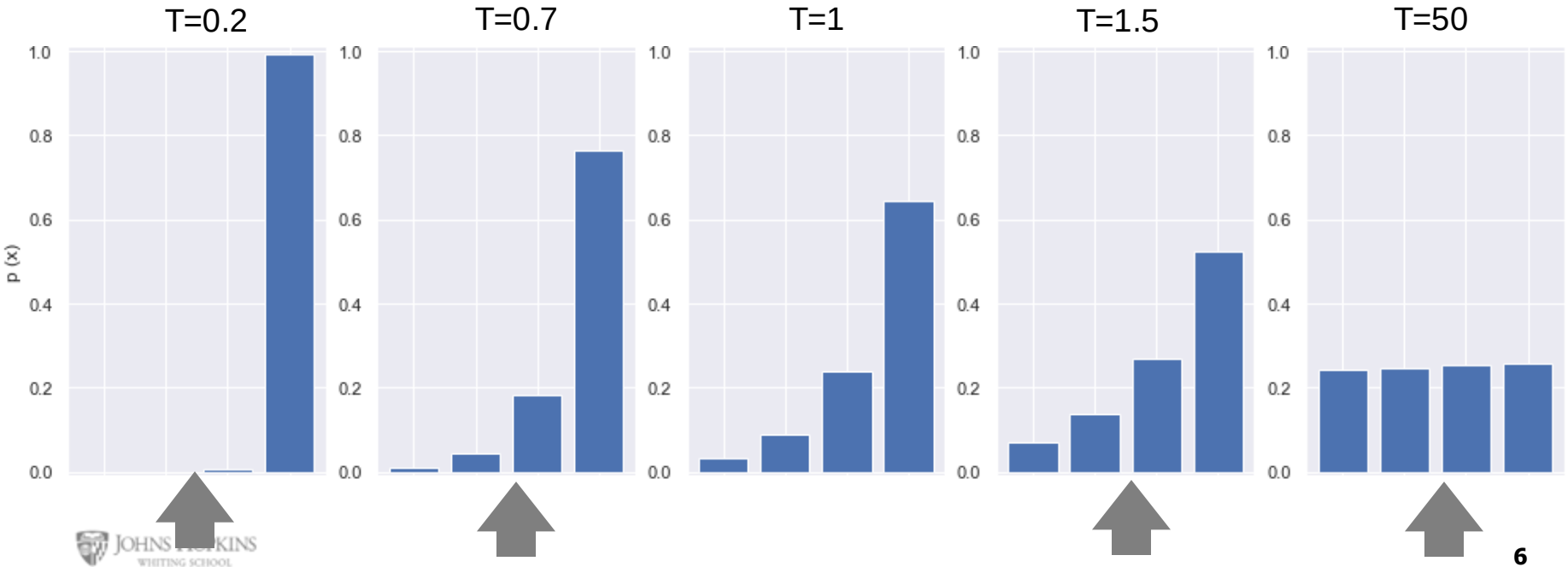
- Let's add parameter T to our Softmax definition:
- Suppose if T=1 (i.e., no Softmax) the dist looks like this:

$$\frac{\exp(z_i/T)}{\sum_j \exp(z_j/T)}$$




- **Question:** How will increasing T it change this distribution?

# Quiz: Guess the Temperature



# Temperature Sampling

- Sample according to multinomial distribution given by **Softmax [ P**(X | $x_{<t}$); $T$ **]**

When $T \to 0$, this is similar to greedy (argmax) decoding.

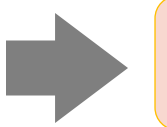
When $T = 1$, this is to sampling from LM distribution.

When $T > 10$, this is to sampling from near uninformative dist.

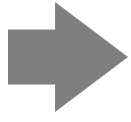
$T \to 0$ $T = 1$ $T = 10$ Temperature

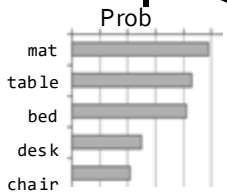
$$x \sim \textbf{Softmax}\,[\,\textbf{P}(X = x|\boldsymbol{x}_{<t});\, T\,]$$

"The cat sat on the [MASK]"



*Some model*

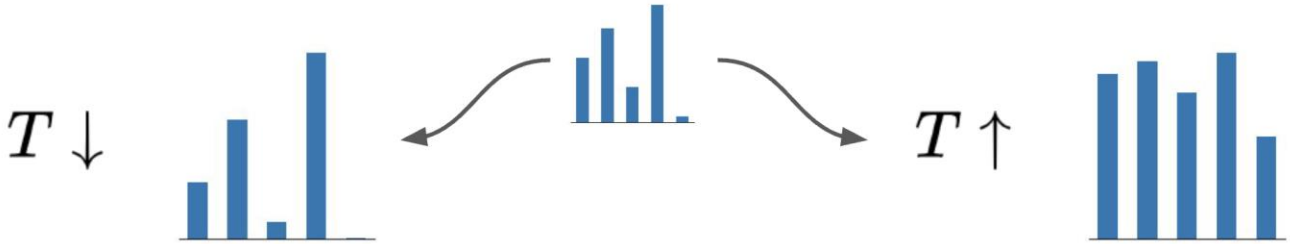







[The Curious Case of Neural Text Degeneration, Holtzman et al., 2020]

# Consequences of Temperature Sampling




One day a cat decided to climb a tree but was caught by a dog. The cat was taken to the vet and the vet said the cat had a broken leg. The cat was then taken to the vet and the vet said the cat had a broken leg and was in a lot of pain. The cat was then .....

One day a cat decided to climb a tree but did not properly rock climb, Zoo workers put her (feet down), disastrously Raphael Staonymusensht October 28, 2014 at 12:38 PM Hot Cat written by .....

Peaky distributions lead to **repetitive** text

Flat distributions lead to **incoherent** text



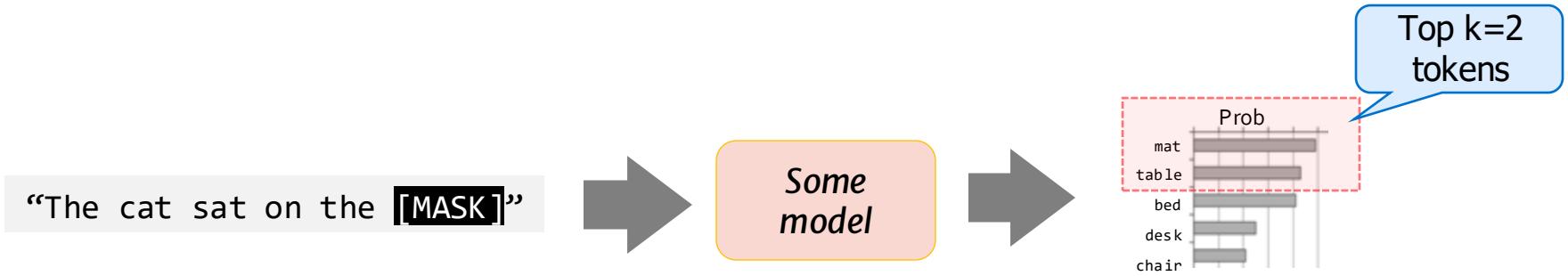
[Slide: Tutorial on "Generating Text from Language Models"]

# Top-k Sampling

- Simply truncate the tail by selecting the $k$ tokens with the largest probability.

$$s(x, \boldsymbol{x}_{<t}) = \begin{cases} \mathbf{P}(X = x | \boldsymbol{x}_{<t}) & \text{if } x \text{ is top}-k \\ 0 & \text{otherwise} \end{cases}$$

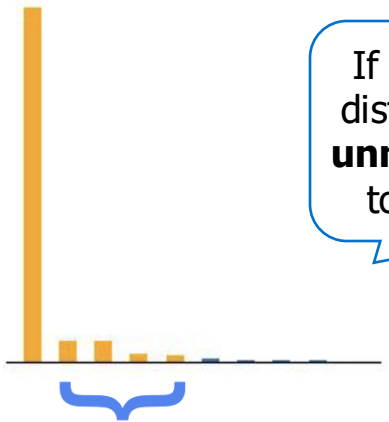
- Sample according to the truncated *score* function $s(x, \boldsymbol{x}_{<t})$ (not a proper probability!)



# Failure of Top-k Sampling

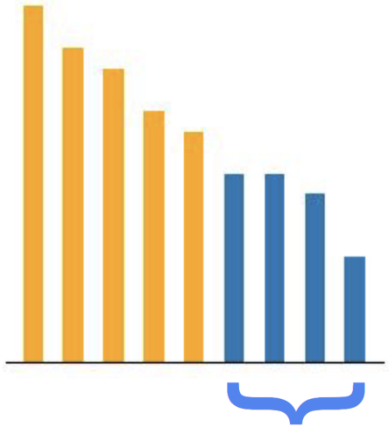
- The use of a fixed $k$ can become problematic across sentences with different entropies.
- Suppose we fix $k = 5$.

$\mathbf{P}(\,.\,|\,\text{“}Hopkins\ is\ located\ in\text{”})$



If the entropy of the distribution is low, we **unnecessarily select** tokens from its tail

$\mathbf{P}(\,.\,|\,\text{“}I\ want\ to\text{”})$



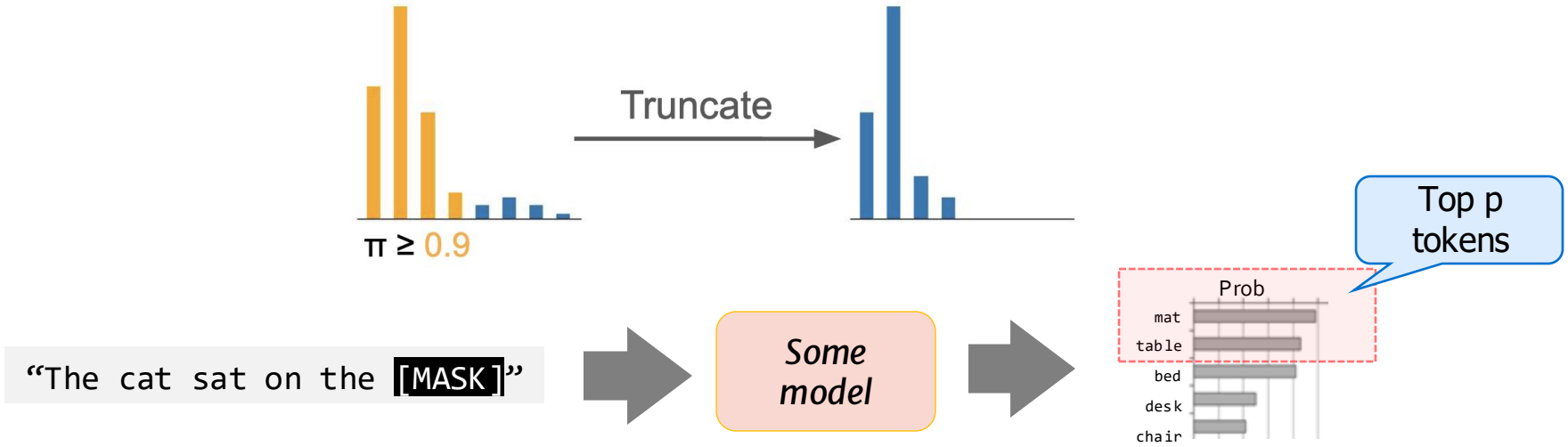
If the entropy of the distribution is high, we **truncate** tokens that are not in the tail.



[Slide: Tutorial on “Generating Text from Language Models”]

# Top-p (Nucleus) Sampling

- Select tail size dynamically by only considering the "nucleus" of the distribution
- On each step, randomly sample from the distribution, but restricted to just the top-p most probable words

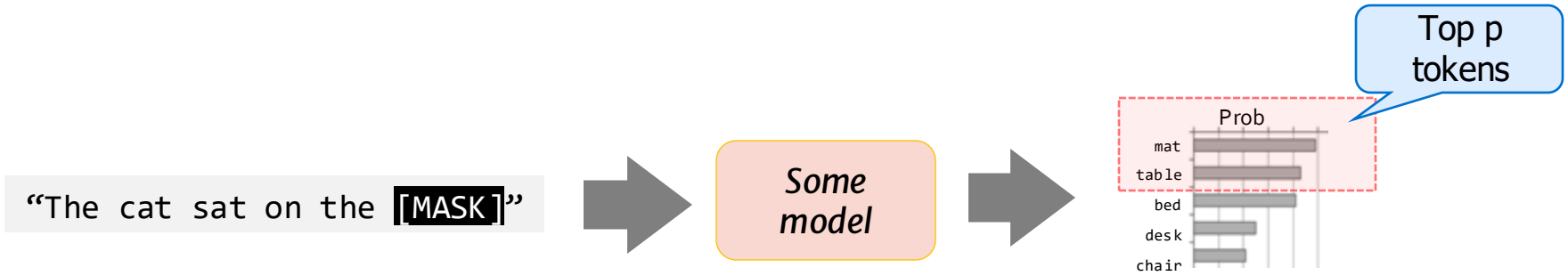



[The Curious Case of Neural Text Degeneration, Holtzman et al., 2020]

# Top-p (Nucleus) Sampling

- Select tail size dynamically by only considering the "nucleus" of the distribution
- On each step, randomly sample from the distribution, but restricted to just the top-p most probable words

$$s(x, \boldsymbol{x}_{<t}) = \begin{cases} \mathbf{P}(X = x|\boldsymbol{x}_{<t}) & \text{if } x \in \underset{\tilde{V} \subset V}{\operatorname{argmin}} |\tilde{V}| \\ 0 & \text{otherwise} \end{cases} \quad s.t. \sum_{x \in \tilde{V}} \mathbf{P}(x|\boldsymbol{x}_{<t}) \geq p$$



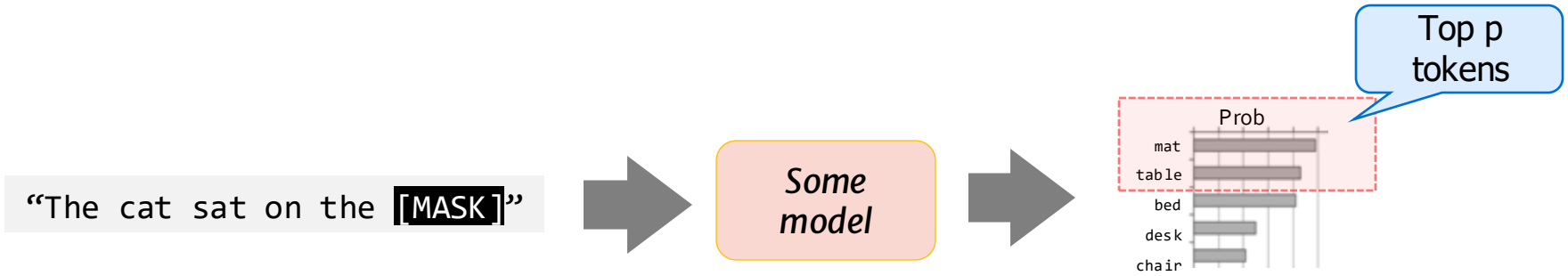



[The Curious Case of Neural Text Degeneration, Holtzman et al., 2020]

# Top-p (Nucleus) Sampling: Quiz

- Select tail size dynamically by only considering the "nucleus" of the distribution
- On each step, randomly sample from the distribution, but restricted to just the top-p most probable words.

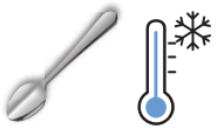
- **Question 1:** If $p = 1$, top-$p$ sampling reduces to what algorithm?
- **Question 2:** How about $p \to 0$?



Local Incoherence

Pure Sampling

The Australian Food Safety Authority has warned Australia's beaches may be revitalised this year because healthy seabirds and seals have been on the move. More than 50,000 seabirds, sea mammals and seahorses have been swept into the sea by the Holden CS118 and Adelaide Airport CS300 from 2013. A major white-bat and umidauda migration across Australia is under way in Australia for the first time, with numbers reaching an estimated 50,000.



Sampling, $t$=0.9

Last week's intense storms and a series of powerful cyclones have been officially blamed for the deaths of at least nine large fin whales near Whitsundays - the largest loss of any species globally. The fin whales: packed in the belly of one killer whale thrashing madly in fear as another tries to bring it to safety. When the colossal animal breached the waters of Whitsundays, he'd been seen tagged for a decade.

Nucleus, $p$=0.95

There has been an unprecedented number of calves caught in the nets of whaling stations that operate in WA. Pilot whales continue to migrate to feeding grounds to feed their calves. They are now vulnerable due to the decline of wild populations; they are restricted to one breeding site each year. Image copyright Yoon Bo Kim But, with sharp decline in wild populations the size of the Petrels are shrinking and dwindling population means there will only be room for a few new fowl.

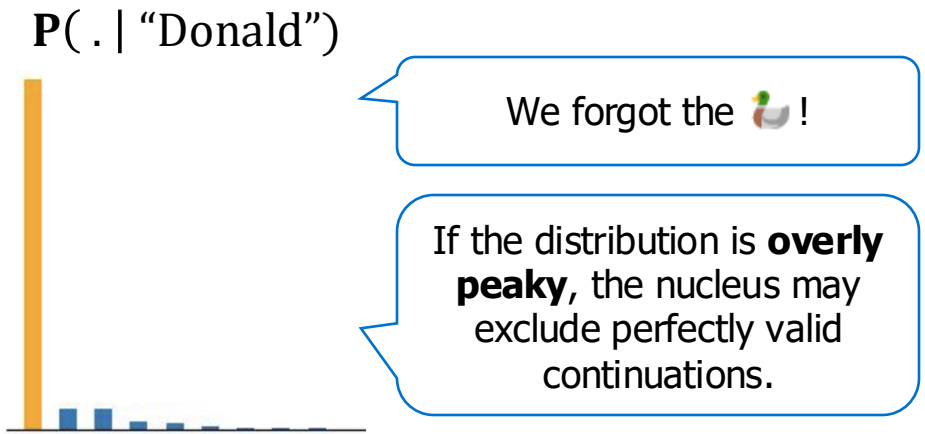
WebText

Poor nutrition has led to a rise in the number of stranded humpback whales on the West Australian coast, veterinary researchers have said. Carly Holyoake, from Murdoch University, at the Australian Veterinary Association's annual conference in Perth on Wednesday, said an unprecedented number of mostly young whales had become stranded on the coast since 2008.

[The Curious Case of Neural Text Degeneration, Holtzman et al., 2020]

# Failure Mode of Top-p Sampling

- Suppose we fix $p = 0.8$. In certain cases, one of the failure modes of top-$k$ sampling can also appear with top-$p$ sampling:



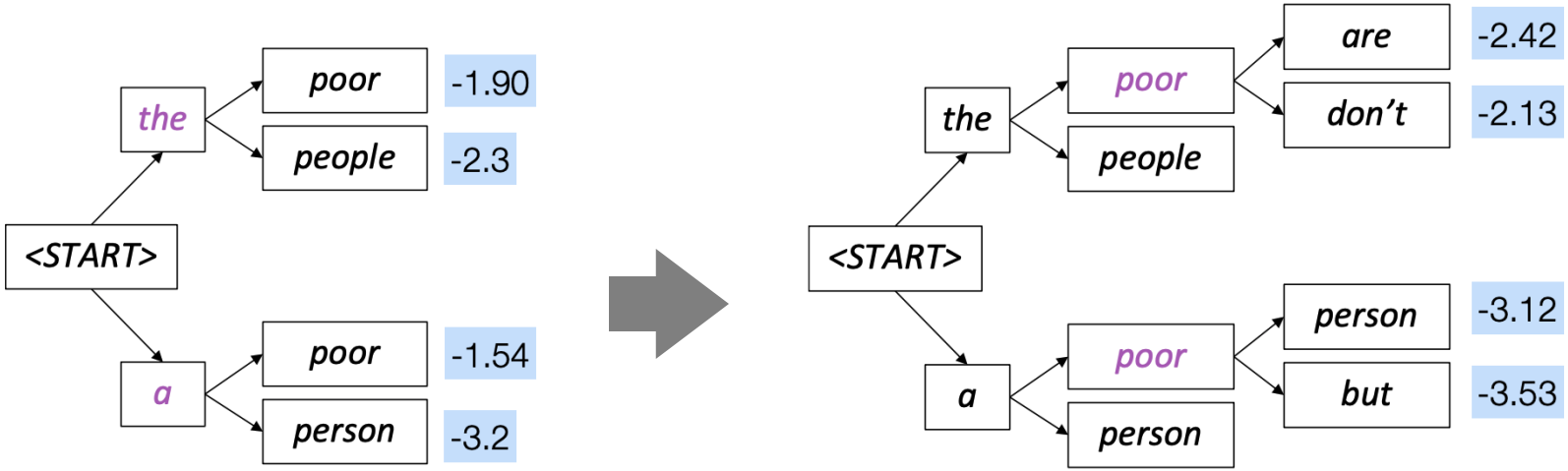



[Slide: Tutorial on “Generating Text from Language Models”]

# Fancier Approaches: Beam Search

- A heuristic search that allows maximizing words probabilities for a window of words
- Take top $k$ continuations, then only keep the $k$ best from the $k^2$, ….

$$X = \text{argmax}_{\tilde{Y} \subset B, |\tilde{Y}| = k} \; \mathbf{P}(\tilde{Y} | \, x_{<t})$$




Details out of scope for us. Feel free to check it in your own time (or take the NLP course!).

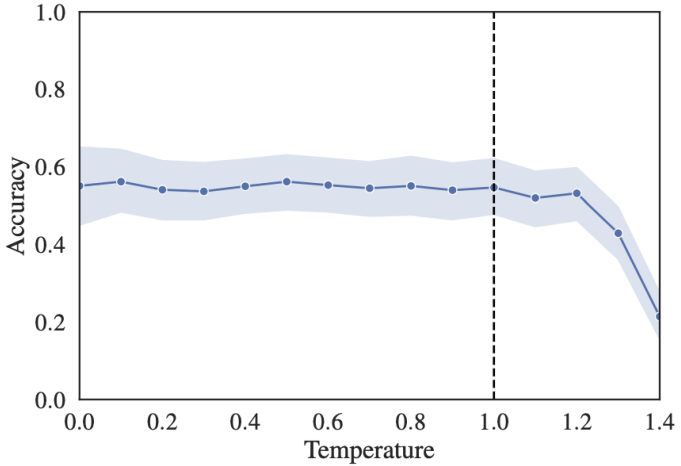
# HuggingFace 🤗 Generation Function

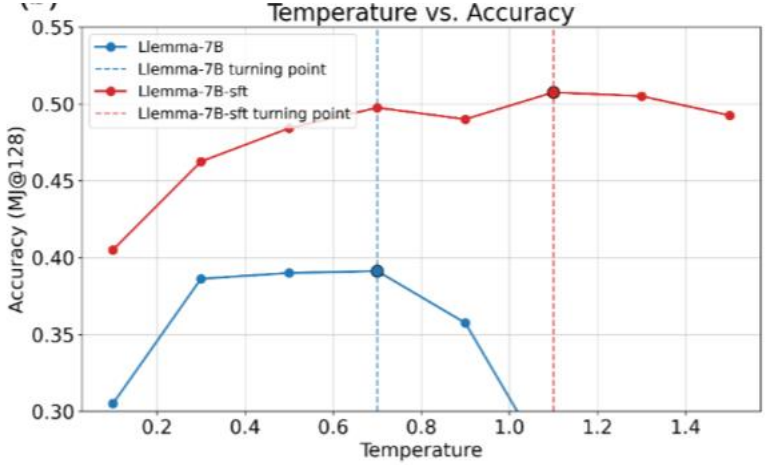
- **min_length** (`int`, *optional*, defaults to 10) — The minimum length of the sequence to be generated.
- **do_sample** (`bool`, *optional*, defaults to `False`) — Whether or not to use sampling ; use greedy decoding otherwise.
- **early_stopping** (`bool`, *optional*, defaults to `False`) — Whether to stop the beam search when at least `num_beams` sentences are finished per batch or not.
- **num_beams** (`int`, *optional*, defaults to 1) — Number of beams for beam search. 1 means no beam search.
- **temperature** (`float`, *optional*, defaults to 1.0) — The value used to module the next token probabilities.
- **top_k** (`int`, *optional*, defaults to 50) — The number of highest probability vocabulary tokens to keep for top-k-filtering.
- **top_p** (`float`, *optional*, defaults to 1.0) — If set to float < 1, only the most probable tokens with probabilities that add up to `top_p` or higher are kept for generation.
- **repetition_penalty** (`float`, *optional*, defaults to 1.0) — The parameter for repetition penalty. 1.0 means no penalty. See this paper for more details.

https://huggingface.co/docs/transformers/main/en/main_classes/text_generation#transformers.GenerationMixin.generate

# Evaluation accuracy vs temperature

- Does higher temperature entail lower accuracy? Not necessarily. But it depends on tasks.
- Figure left: GPT-3.5 results on a multiple-choice questions w/ CoT prompting (coming up).
- Figure right: Llemma-7B on math problems.



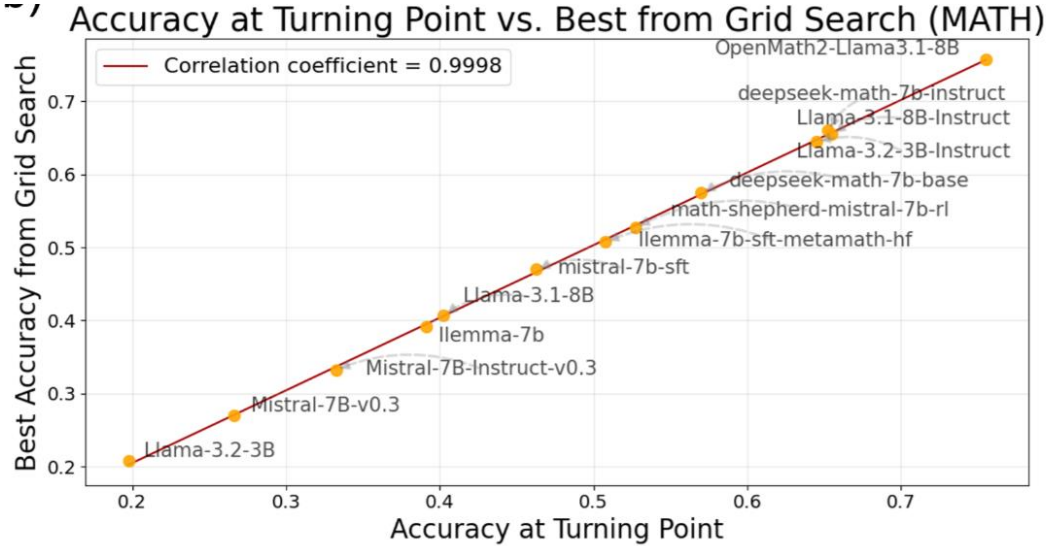






The Effect of Sampling Temperature on Problem Solving in Large Language Models, 2024
Optimizing Temperature for Language Models with Multi-Sample Inference, 2025

# Evaluation accuracy vs temperature

- Does higher temperature entail lower accuracy? It also depends on models.
- Figure: temperature that leads to the highest accuracy with different models.




Optimizing Temperature for Language Models with Multi-Sample Inference, 2025

# Summary on Sampling Algorithms

- Greedy decoding: a simple method; gives low-surprising output
- Sampling methods are a way to balance diversity vs quality
  - Temperature is a way of balancing for distribution peakiness
  - Top-k and top-p sampling allows you to control diversity

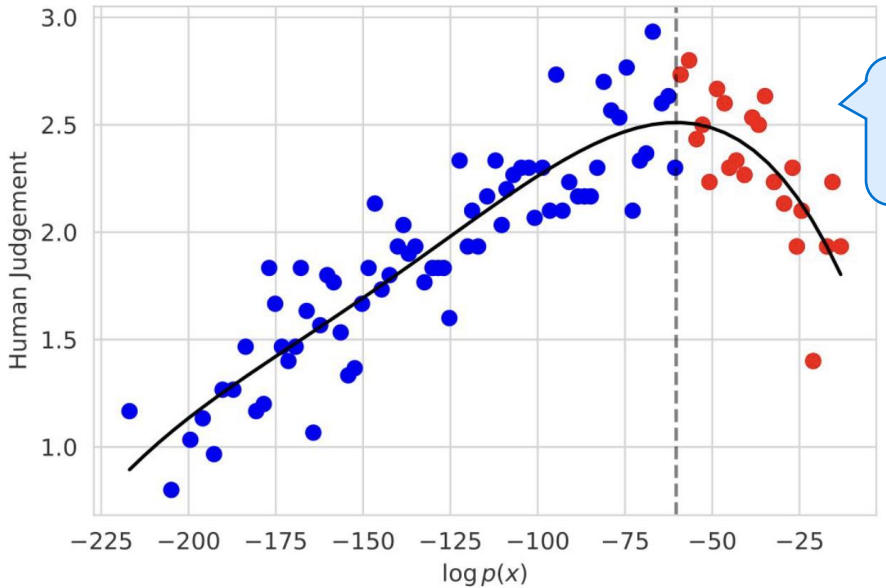
# Many others that we do not cover …

- There are many other algorithms for sampling sentences from LMs that we will not see in this course.
  - Mirostat: https://arxiv.org/abs/2007.14966
  - Eta-sampling and Epsilon-sampling: https://aclanthology.org/2022.findings-emnlp.249/
  - …

# Bonus Content: More Sampling Algorithms

# Perplexity and Generation Quality

- Surprisal: $-\log \mathbf{P}(x|\boldsymbol{x}_{<t})$




High-quality text (human judgement) has been shown to exhibit surprisal values that fall within a narrow range

This motivates decoding algorithms that maintain a target "surprisal" during generation.



Trading Off Diversity and Quality in Natural Language Generation, 2021

[Slide: Tutorial on "Generating Text from Language Models"]

# Mirostat

- **Idea:** Perform top-$k$ sampling; Tune $k$ so that the surprisal of the generated text is close to a target value $\tau$.
- At each generation step:
  - Choose the top-k set to sample a token with a desirable surprisal:

$$k = |x \in V : -\log \mathbf{P}(x|\boldsymbol{x}_{<t}) \leq 2\tau\,|$$

  - Update using the surprisal of the sampled token $x^*$

$$\tau = \tau - \eta(-\log \mathbf{P}(x^*|\boldsymbol{x}_{<t}) - \tau)$$

- The dynamic value of $k$ helps avoiding repetitions and incoherent text.

Typical of small $k$

Typical of large $k$



Mirostat: A Neural Text Decoding Algorithm that Directly Controls Perplexity, 2021

[Slide: Tutorial on "Generating Text from Language Models"]

# Typical Sets

- In information theory, is a set of sequences whose prob is close to a "typical" value.

$$2^{-n(H(X)+\varepsilon)} \leqslant p(x_1, x_2, \ldots, x_n) \leqslant 2^{-n(H(X)-\varepsilon)}$$

- where: $H(X) = -\sum_{x \in \mathcal{X}} p(x) \log_2 p(x)$

- Taking the log of all sides:

$$H(X) - \varepsilon \leq -\frac{1}{n} \log_2 p(x_1, x_2, \ldots, x_n) \leq H(X) + \varepsilon.$$

# Typical Decoding

- Select the set of tokens that are locally "typical".

$$s(x, \boldsymbol{x}_{<t}) = \begin{cases} \mathbf{P}(X = x|\boldsymbol{x}_{<t}) & \text{if } x \in \underset{\widetilde{V} \subset V}{\operatorname{argmin}} \left| H\left(\mathbf{P}(.|\boldsymbol{x}_{<t})\right) + \log \mathbf{P}(x|\boldsymbol{x}_{<t}) \right| \\ & s.t. \sum_{x \in \widetilde{V}} \mathbf{P}(x|\boldsymbol{x}_{<t}) \geq p \\ 0 & \text{otherwise} \end{cases}$$

- Based on the assumption that natural language has an information content close to the expected information content given prior context.
- Locally typical sampling does not necessarily select high-probability tokens, avoiding repeated or dull text and promoting lexical diversity.



Locally Typical Sampling, 2023

[Slide: Tutorial on "Generating Text from Language Models"]

# $\eta$-Sampling

- Enforces the concurrent application of two principles:
  1. **Absolute probability principle:** The scoring function shouldn't truncate high-probability tokens above a small threshold.
  2. **Relative probability principle:** The scoring function should only truncate low-probability words if they are low relative to the rest of the distribution.

$$s(x, \boldsymbol{x}_{<t}) = \begin{cases} \mathbf{P}(X = x|\boldsymbol{x}_{<t}) & \text{if } x \in \{y \in V |\, \mathbf{P}(x|\boldsymbol{x}_{<t}) > \eta\} \\ & s.t.\ \eta = \min(\epsilon, \sqrt{\epsilon}\exp(-H(.|\boldsymbol{x}_{<t}))) \\ 0 & \text{otherwise} \end{cases}$$



Truncation Sampling as Language Model Desmoothing, 2022

[Slide: Tutorial on "Generating Text from Language Models"]