

# *Applied Deep Learning*



## NLG Decoding



September 25th, 2024

<http://adl.miulab.tw>



National  
Taiwan  
University  
國立臺灣大學

- NLG Review
  - Language Modeling
  - Conditional Language Modeling
- Decoding Algorithm
  - Greedy
  - Beam Search
  - Sampling
  - Top- $k$  Sampling
  - Nucleus Sampling

# Natural Language Generation

---

- Many tasks contain NLG
  - Machine Translation
  - Abstractive Summarization
  - Dialogue Generation
  - Image Captioning
  - Creative Writing
    - Storytelling, poetry generation
  - ...

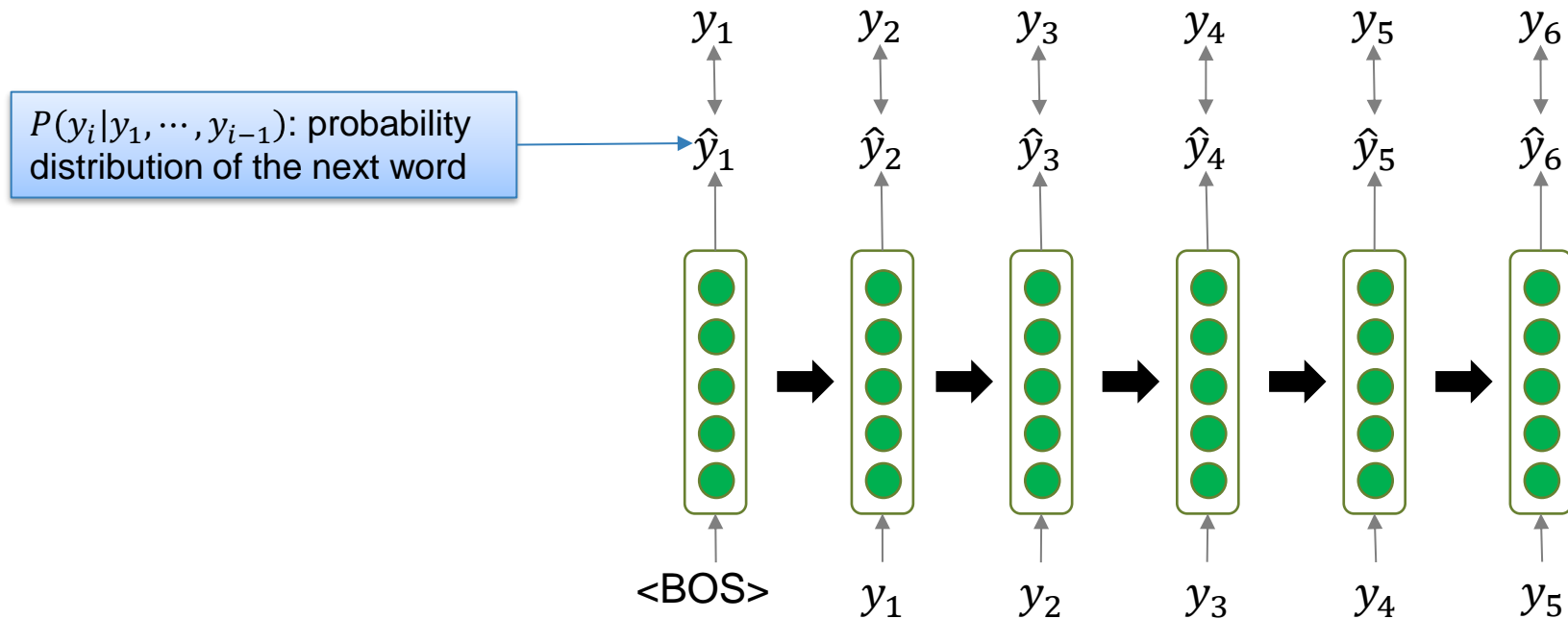
# Language Modeling

- Goal: predicting the next word given the words so far

$$P(y_i | y_1, \dots, y_{i-1})$$

- Language model** is to estimate the probability distribution
  - RNN-LM uses RNN for modeling the distribution
  - GPT uses Transformer for modeling the distribution

# Language Modeling



Idea: pass the information from the previous hidden layer to leverage all contexts

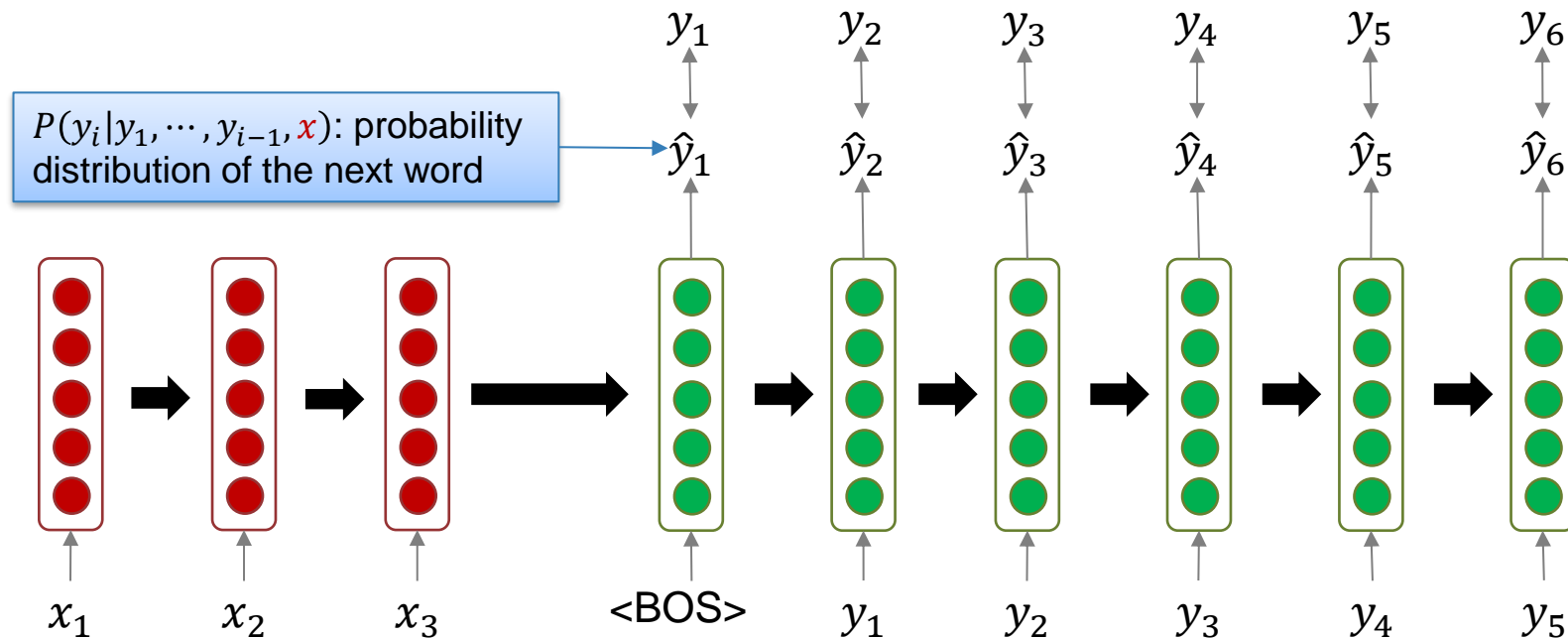
# Conditional Language Modeling

- Goal: predicting the next word given the words so far, and other input  $x$

$$P(y_i | y_1, \dots, y_{i-1}, x)$$

- Conditional language modeling tasks
  - Machine translation ( $x$  = source sentence,  $y$  = target sentence)
  - Summarization ( $x$  = document,  $y$  = summary)
  - Dialogue ( $x$  = dialogue context,  $y$  = response)
  - Image captioning ( $x$  = image,  $y$  = caption)
  - ...

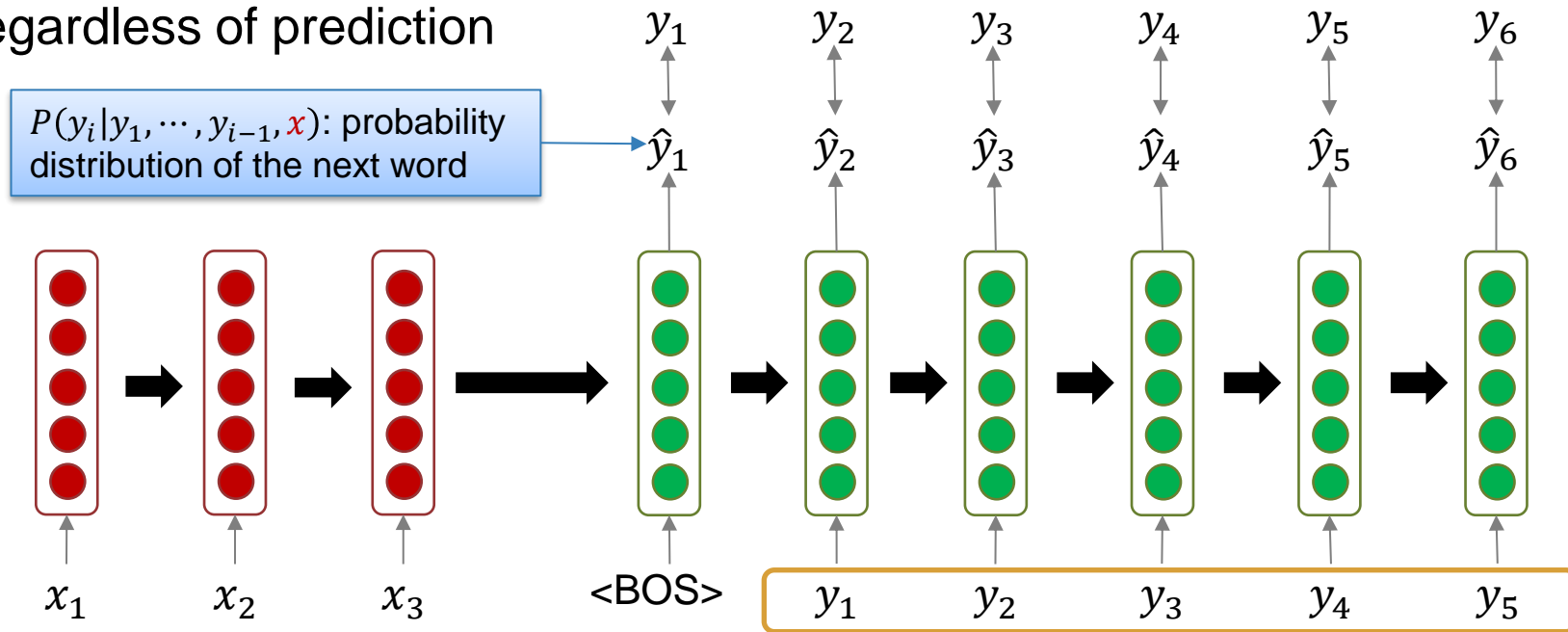
# Conditional Language Modeling



An encoder-decoder model or a decoder only architecture can condition on context

# Teacher Forcing

- During training, feeding the **gold target** sentence into the decoder regardless of prediction



Issue: mismatch between training and testing




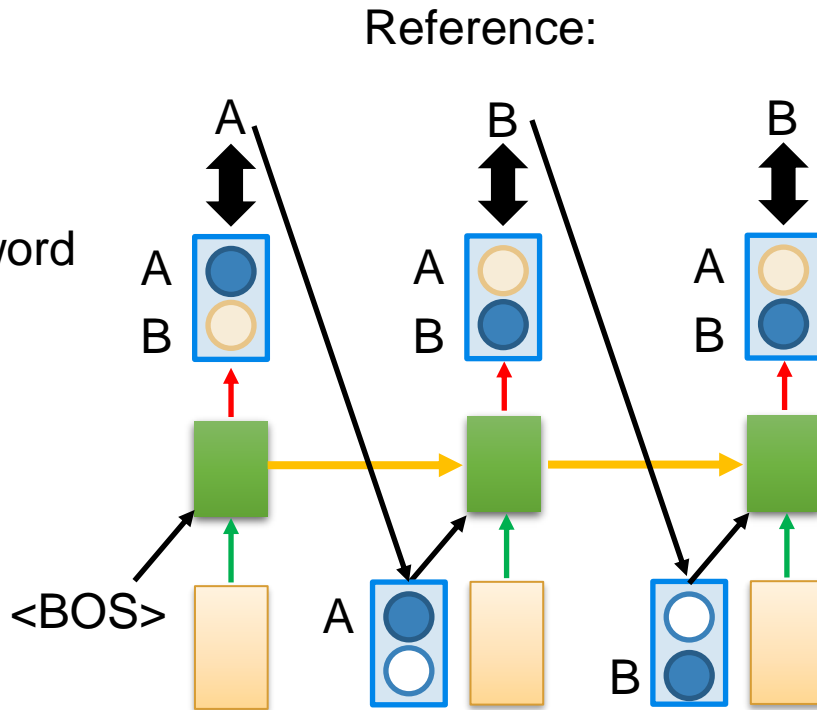
# Mismatch between Train and Test

## Training

$$C = \sum_t C_t$$

minimizing cross-entropy of each word

 : condition

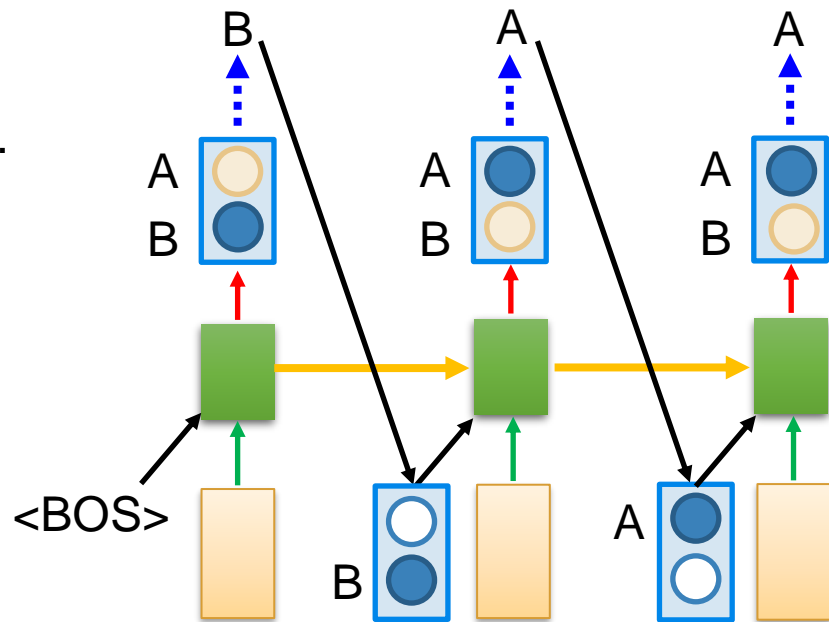


# Mismatch between Train and Test

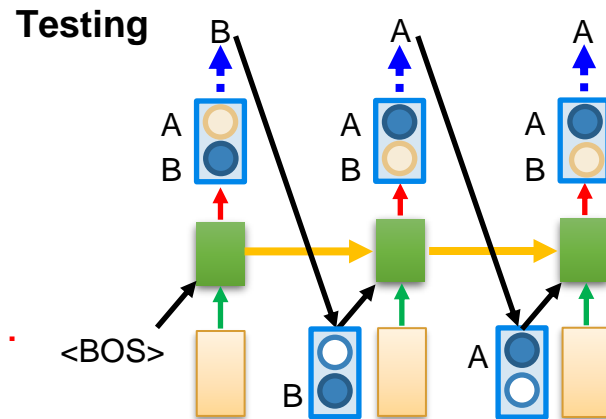
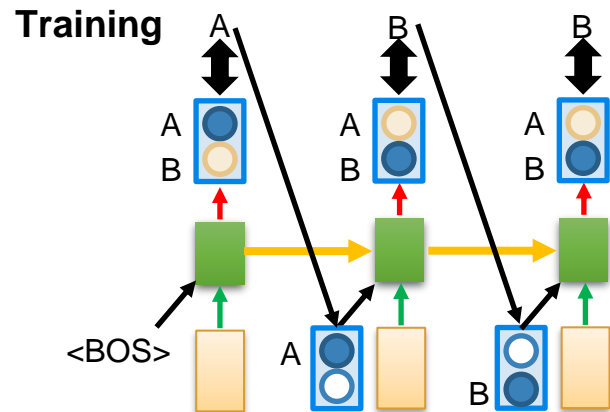
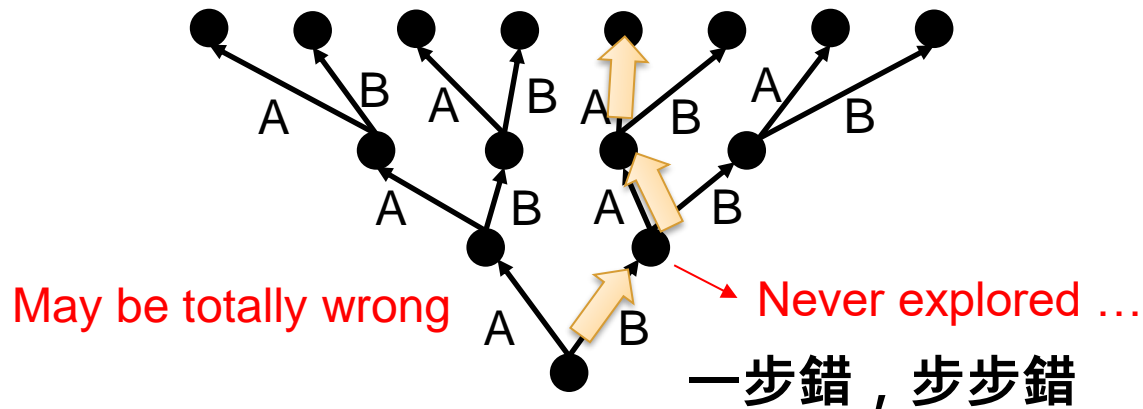
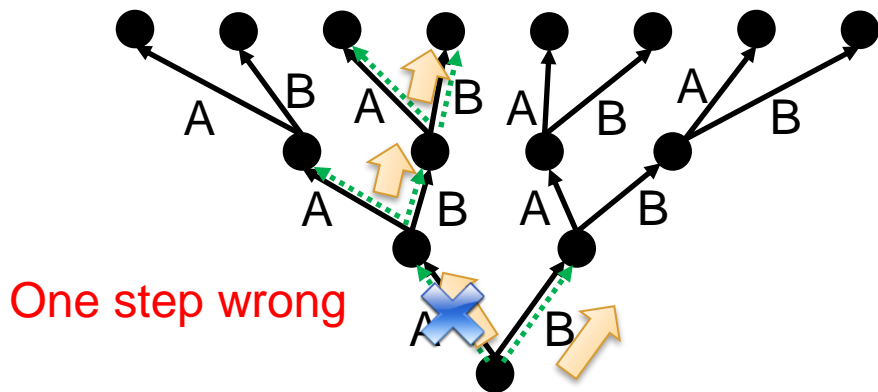
## Generation

- Testing: Output of model is the input of the next step.
  - Reference is unknown
- Training: the inputs are reference.

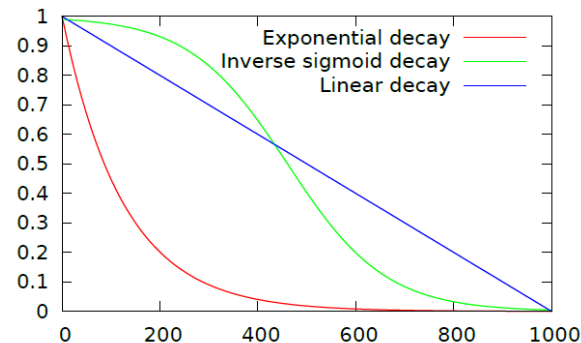
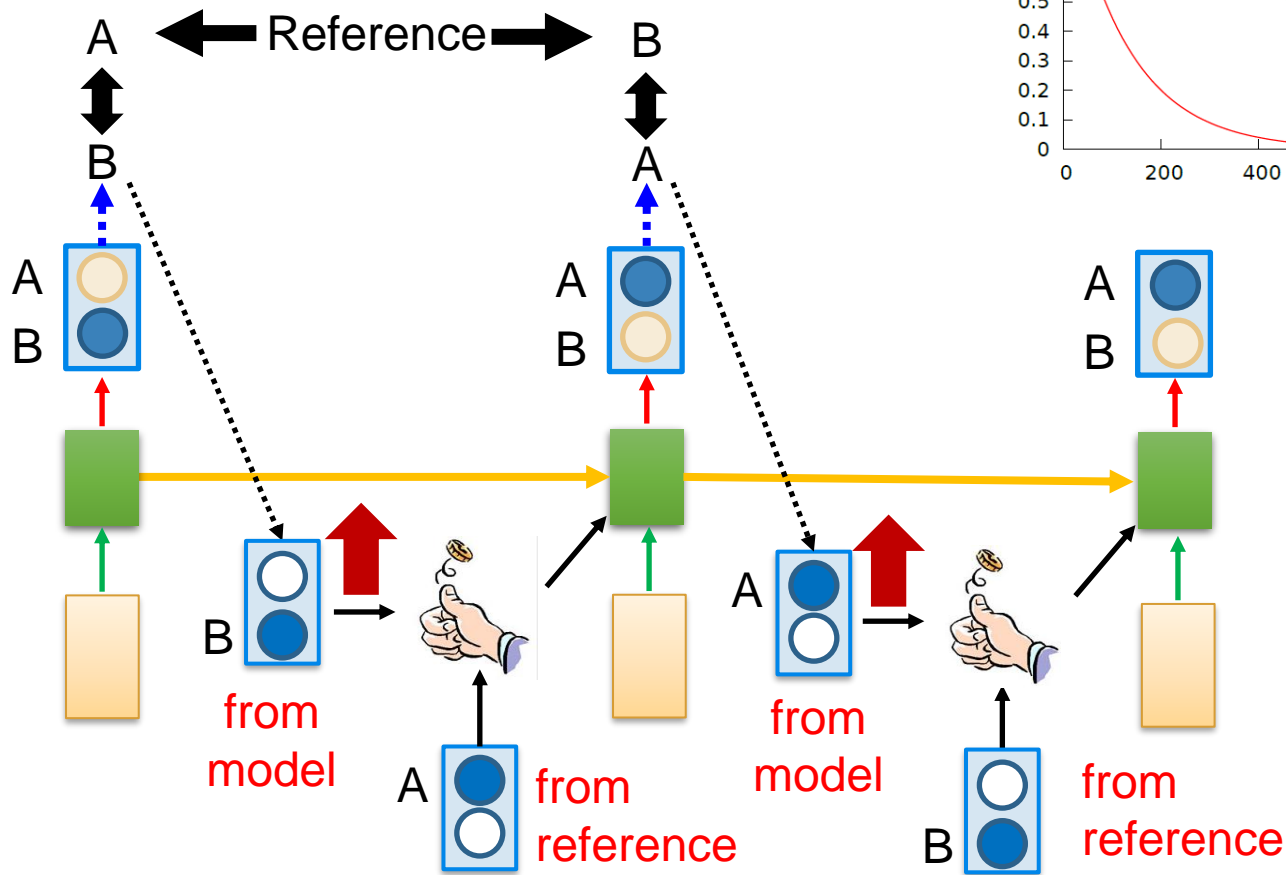
Exposure Bias



# Exposure Bias



# Scheduled Sampling



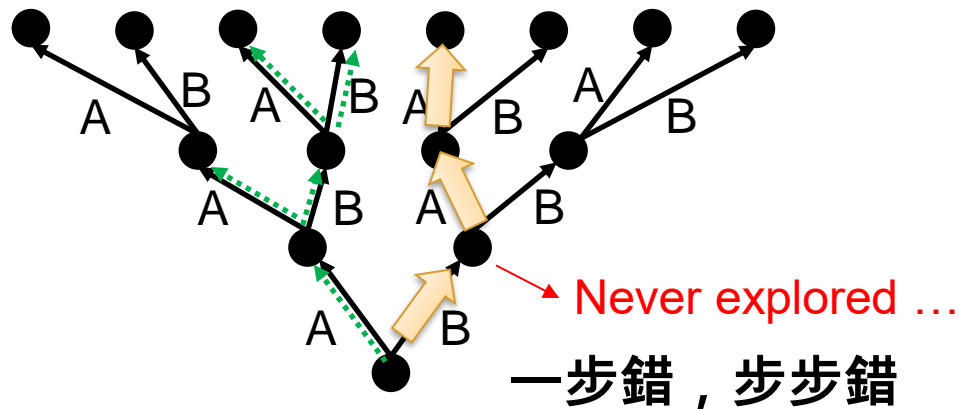
# Scheduled Sampling

## Image captioning on MSCOCO

	<b>BLEU-4</b>	<b>METEOR</b>	<b>CIDER</b>
Always from reference	28.8	24.2	89.5
Always from model	11.2	15.7	49.7
Scheduled Sampling	30.6	24.3	92.1

# No Scheduled Sampling in LLM Training

## Exposure bias



LLM pre-training may explore much more paths from large data

15

# Decoding Algorithm

Strategy of Word Generation

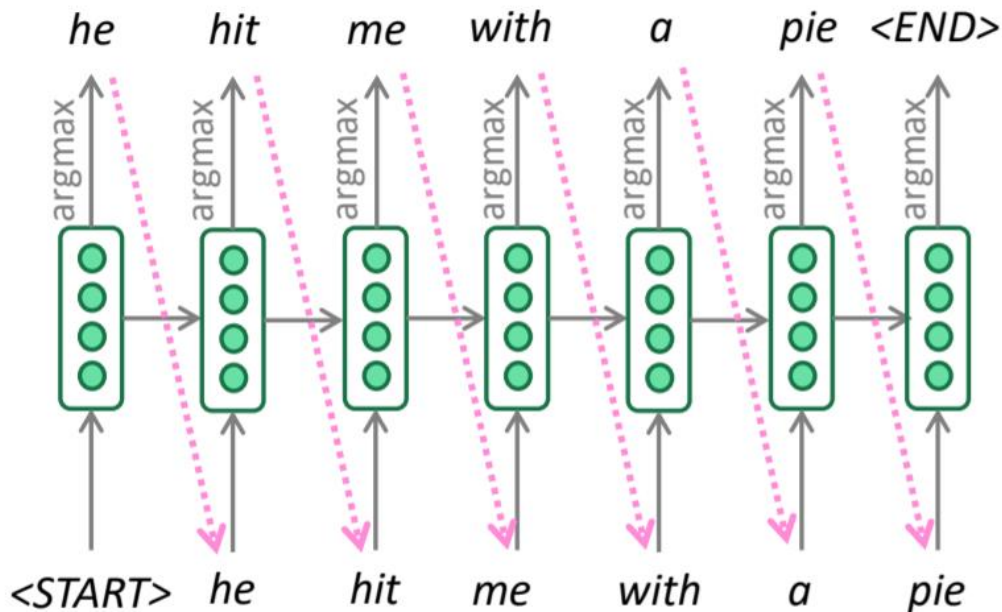
# Decoding Algorithm

- With a trained (conditional) LM, a decoding algorithm decides how to generate texts from the LM.
- Decoding Algorithms
  - Greedy
  - Beam Search
  - Sampling
  - Top- $k$  Sampling
  - Nucleus Sampling



# Greedy

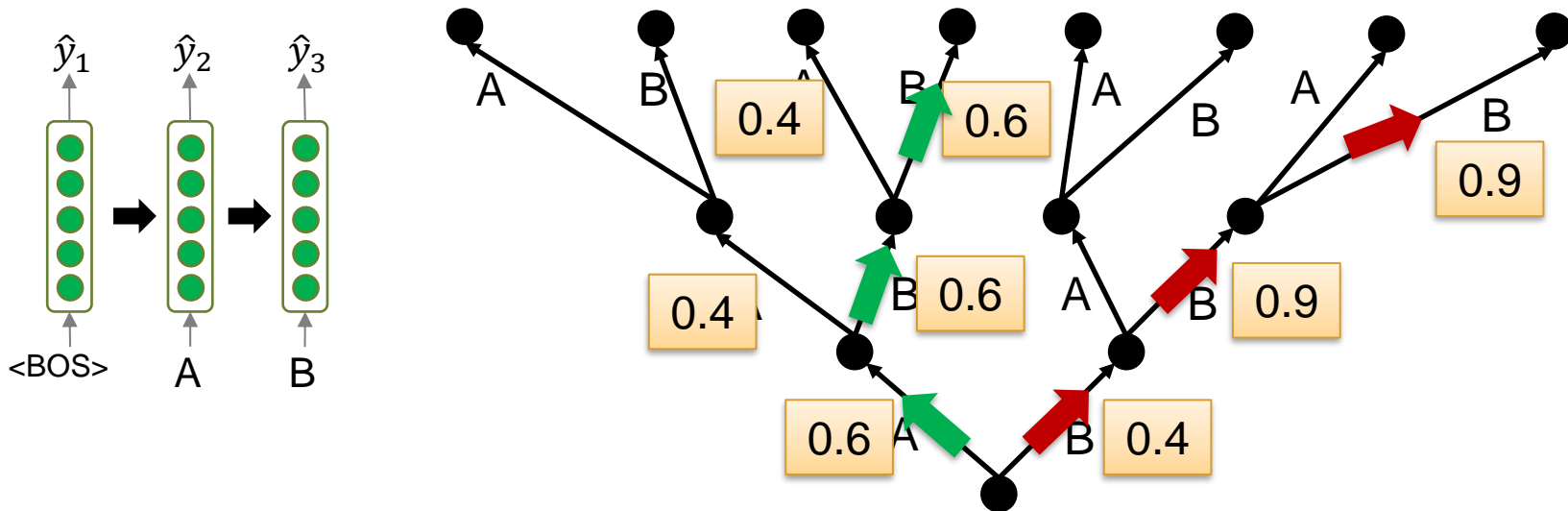
- Strategy: choosing **the most probable** word (argmax)



Output can be poor due to lack of backtracking

# Suboptimal Issue

- Unexplored path may have higher probability.



The **red** path has higher score.

Issue: Impossible to check all paths

# Greedy Example

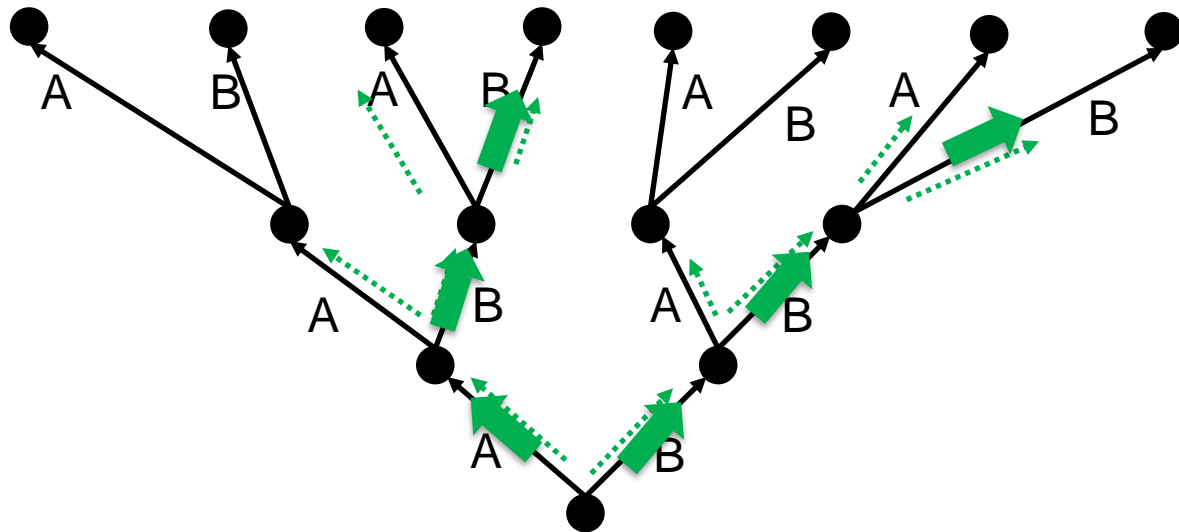
**Initial:** Dwight arose from his bed. He walked down stairs, He made his breakfast, and he sat at the finely crafted wooden dinner table. At his right, a cup of coffee. At his left, the news paper. The crossword puzzle was particularly interesting.

**Continuation:** The headline read: "The New York Times." The headline read: "The New York Times." The headline said: "The New York Times."

# Beam Search

- Strategy: keeping track of the ***k* most probable** sequences and finding a better one

Keep several best paths at each step (beam size = 2)



# Effect of Beam Size

- Small beam size
  - Ungrammatical, unnatural, incorrect, etc.
- Large beam size
  - Reduce some above issues
  - Computationally expensive
  - Introduce other issues
    - Chit-chat dialogues with large beam often generate generic sentences

# Effect of Beam Size in Chit-Chat Dialogues

I mostly eat a fresh and raw diet, so I save on groceries



Beam Size	Model Response
1	<i>I love to eat healthy and eat healthy</i>
2	<i>That is a good thing to have</i>
3	<i>I am a nurse so I do not eat raw food</i>
4	<i>I am a nurse so I am a nurse</i>
5	<i>Do you have any hobbies?</i>
6	<i>What do you do for a living?</i>
7	<i>What do you do for a living?</i>
8	<i>What do you do for a living?</i>

**Small Beam Size:**  
More on-topic but nonsensical; bad English

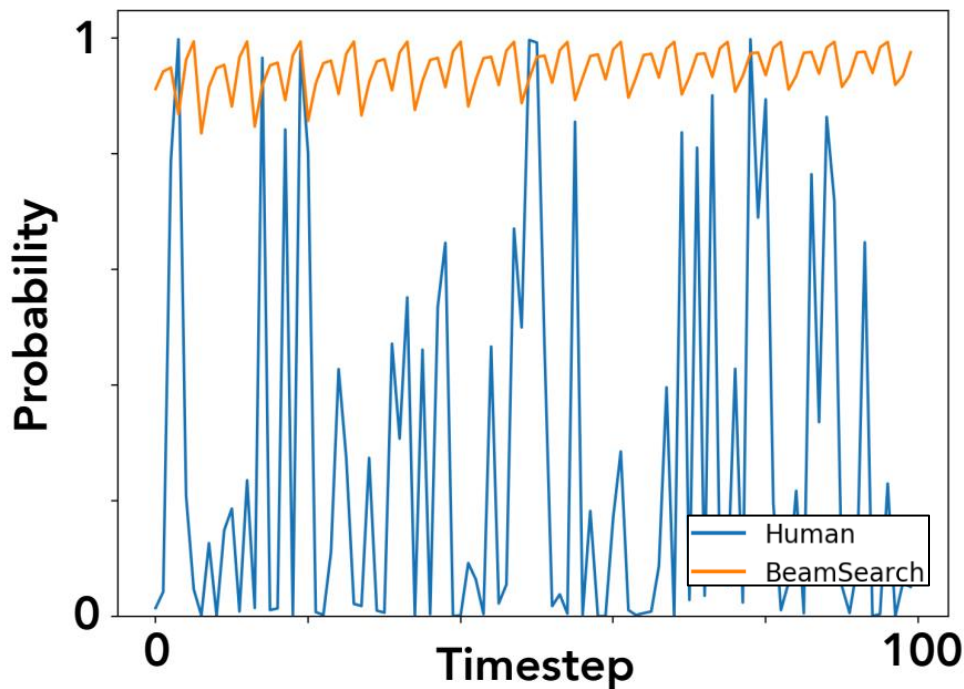
**Large Beam Size:**  
safe, “correct” response, but generic and less relevant

Finding a proper beam size is not trivial



# Distribution Difference

- ⦿ The natural distribution of human text has lots of **spikes**.
- ⦿ In contrast, the distribution of machine text is **high** and **flat**!





# Why Doesn't Maximization Work

- Successful language models all rely heavily on attention, which easily learns to **amplify a bias** towards *repetition*.
- Maximization is problematic in **high-entropy** timesteps, regardless of the quality of the language model.
- Humans aren't attempting to maximize probability, they're trying to achieve goals. (Goodman, 2016)

# Sampling-Based Decoding

- Strategy: choosing the next word with randomness (from a distribution)
- Sampling
  - Randomly sample the word via the **probability distribution** instead of argmax

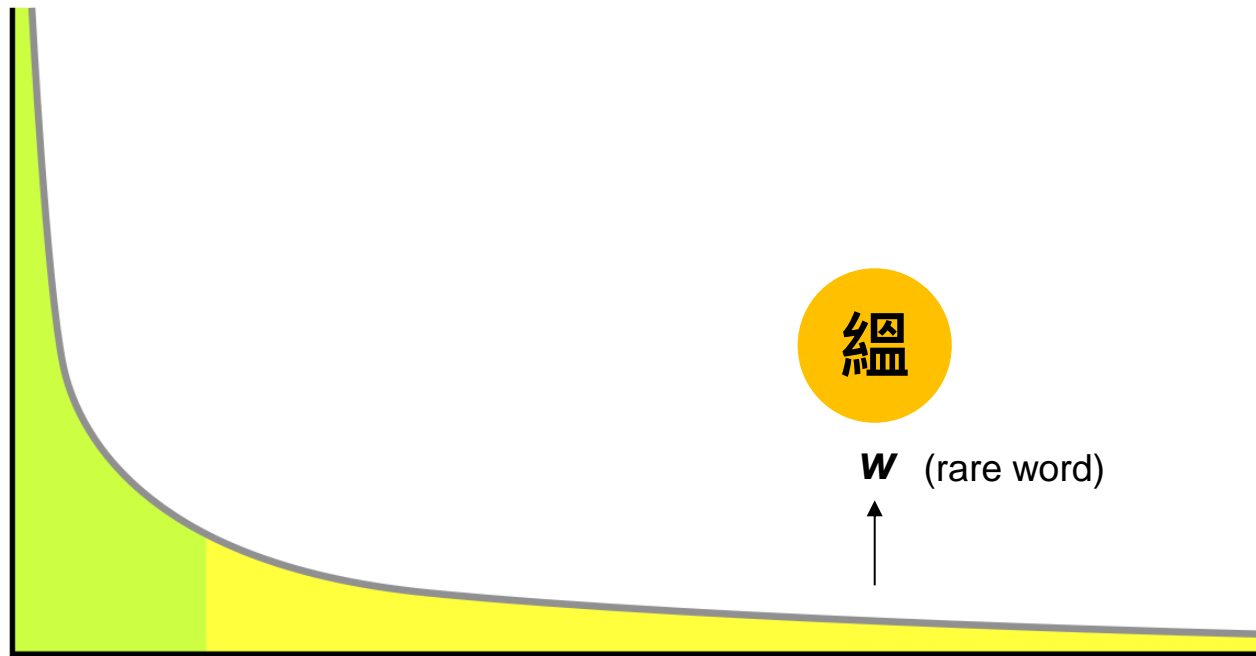
# Sampling Example

**Initial:** Dwight arose from his bed. He walked down stairs, He made his breakfast, and he sat at the finely crafted wooden dinner table. At his right, a cup of coffee. At his left, the news paper. The crossword puzzle was particularly interesting.

**Continuation:** He had opened the crossword puzzle and was pointing the newspaper from it. And the title: 12:50pm how happy has white rabbit been? why is They declining white rabbit?

The *(long) tail* of the distribution is where the quality of LMs become worse.

# Issue of Long Tail Distribution



# Sampling-Based Decoding

- Strategy: choosing the next word with randomness (from a distribution)
- Sampling
  - Randomly sample the word via the **probability distribution** instead of argmax
- Top- $k$  Sampling
  - Sample the word via distribution but **restricted to the top- $k$**  probable words
  - $k=1$  is greedy,  $k=V$  is pure sampling
  - Increasing  $k$  gets more diverse / risky output
  - Decreasing  $k$  gets more generic / safe output

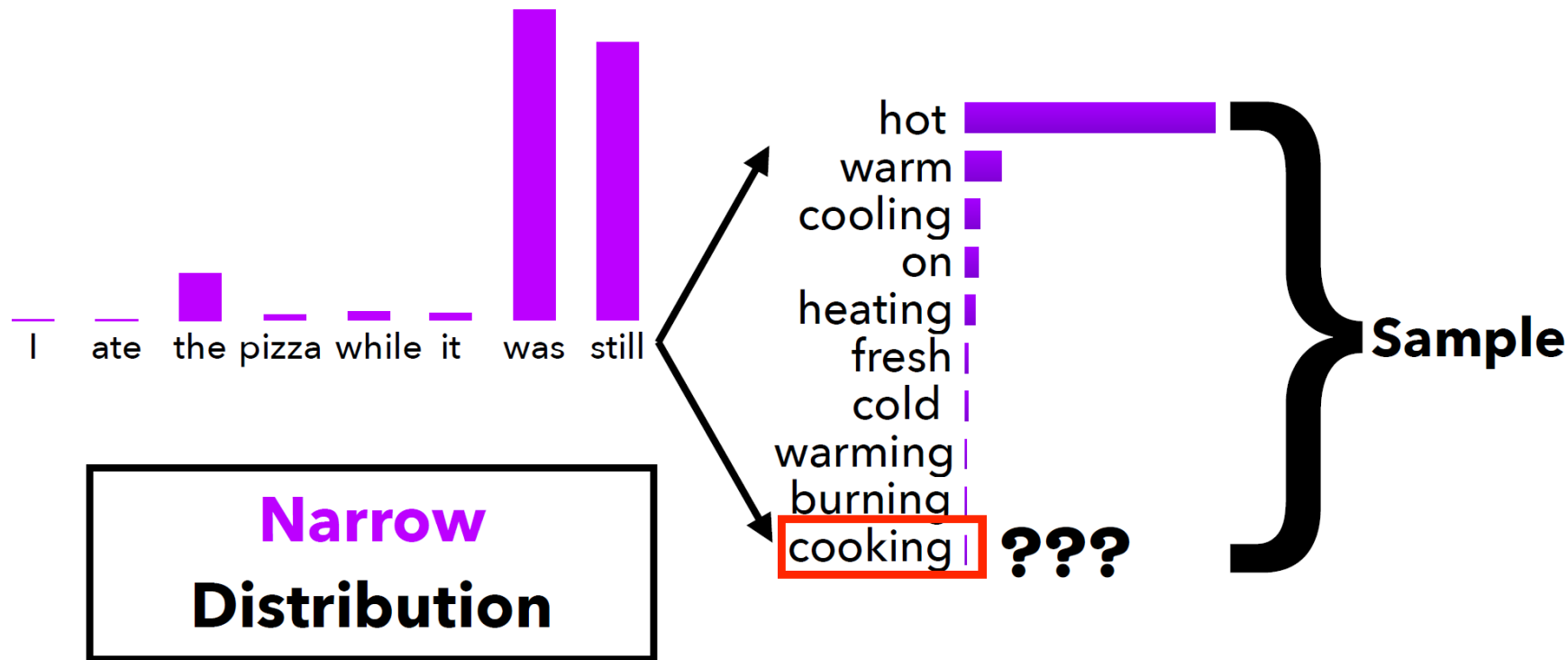
Balancing between diversity and safety is an important direction

# Top- $k$ Sampling Example

**Initial:** Dwight arose from his bed. He walked down stairs, He made his breakfast, and he sat at the finely crafted wooden dinner table. At his right, a cup of coffee. At his left, the news paper. The crossword puzzle was particularly interesting.

**Continuation:** He had seen the news, but had not read the New York times or the times. The local post would have been much quicker, perhaps even better.

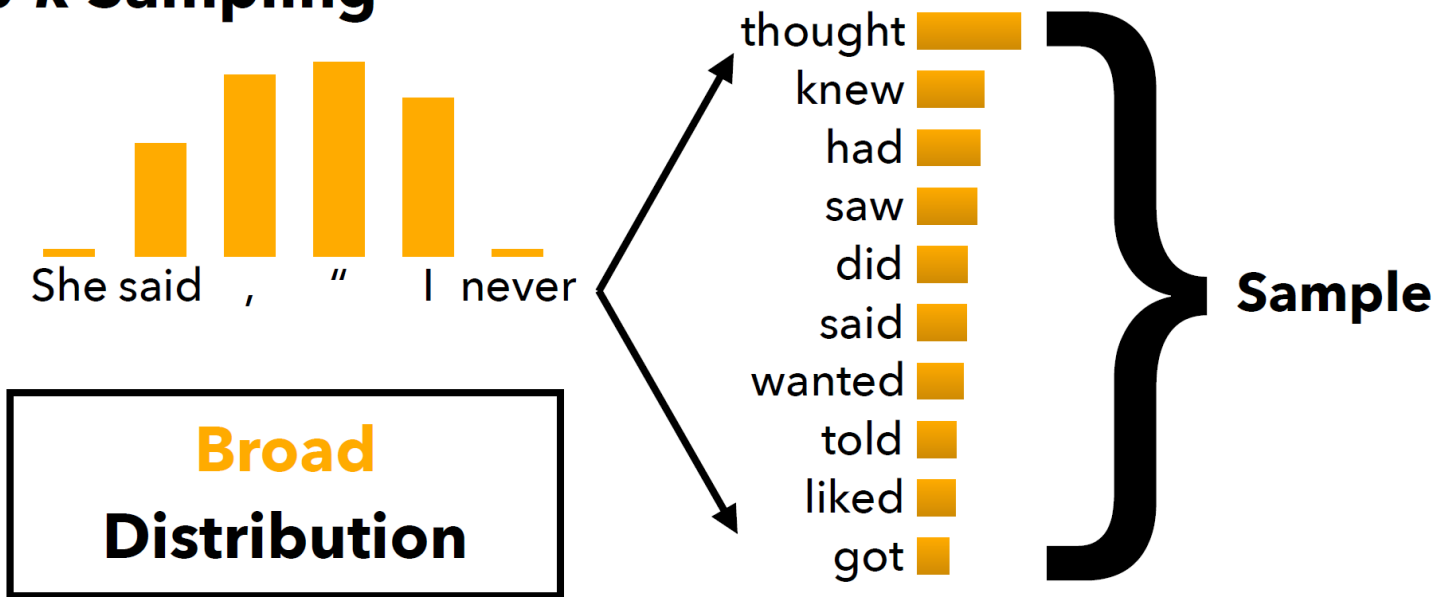
# Top- $k$ Issue 1: Narrow Distribution



High confidence  $\rightarrow$  some extremely low probability choices

# Top- $k$ Issue 2: Broad Distribution

## Top- $k$ Sampling



Low confidence → generic choices



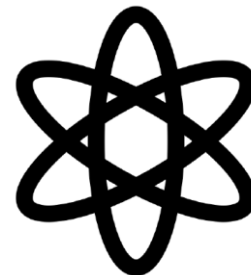
# Nucleus (Top- $p$ ) Sampling

- Sampling from a subset of vocabulary with the most probability mass

$$w_i \sim V^{(p)}$$

where

$$V^{(p)} = \sup_{V' \subset V} \sum_{x \in V'} P(x|w_1 \cdots w_{i-1}) \geq p$$



Nucleus sampling can *dynamically* shrinking and expanding top-k.

# Nucleus Sampling Example

**Initial:** Dwight arose from his bed. He walked down stairs, He made his breakfast, and he sat at the finely crafted wooden dinner table. At his right, a cup of coffee. At his left, the news paper. The crossword puzzle was particularly interesting.

**Continuation:** It was on the ground floor of the Imperial Hotel. He could hear the TV from the lobby of the palace. There were headlines that would make a cop blush.

35

# Generation Controlling

Encourage what we want and penalize what we don't want

# Temperature

## 1. Softmax

$$P(w) = \frac{\exp(s_w)}{\sum_{w' \in V} \exp(s_{w'})}$$

softmax: LM computes a prob dist by applying softmax to a vector of scores

## 2. Softmax temperature: applying a **temperature hyperparameter** $\tau$ to the softmax

$$P(w) = \frac{\exp(s_w / \tau)}{\sum_{w' \in V} \exp(s_{w'} / \tau)}$$

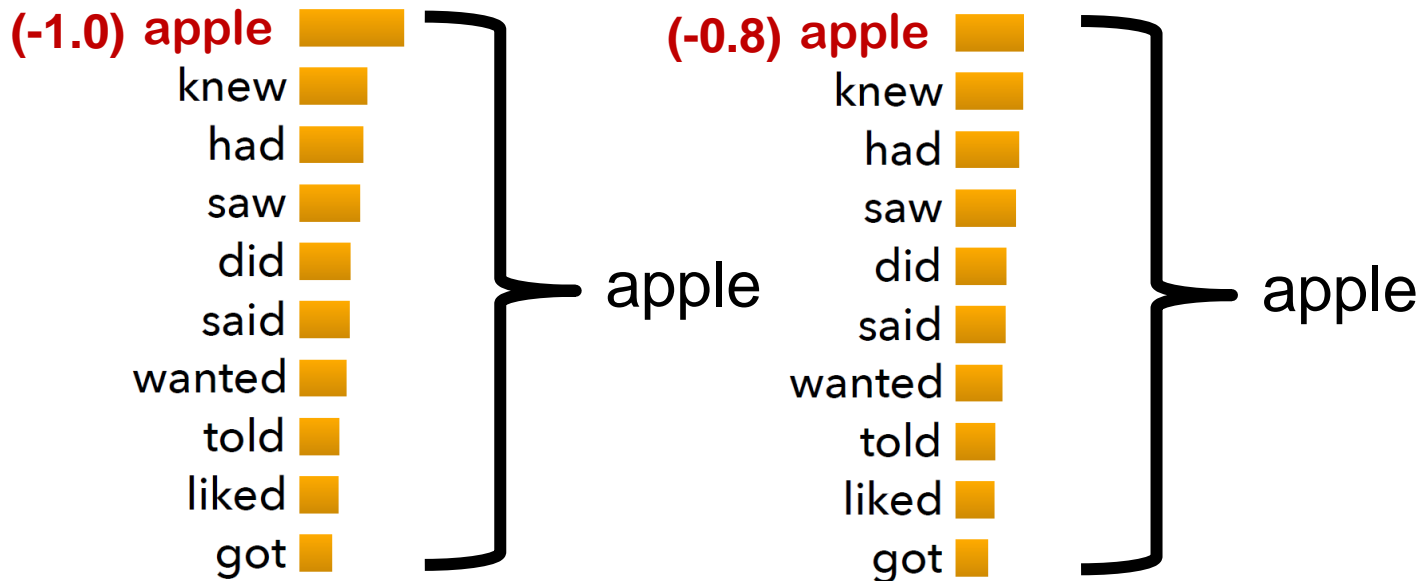
- Higher temperature:  $P(w)$  becomes more uniform  $\rightarrow$  more diversity
- Lower temperature:  $P(w)$  becomes more spiky  $\rightarrow$  less diversity

softmax temperature is not a decoding algorithm, which is the way of **controlling the diversity** during testing via any decoding algorithm

# Repetition Penalty

- Idea: discourage repetitions

$$\log P(w) \leftarrow \log P(w) + \text{penalty}$$



## 38 Frequency / Presence Penalty

### ◎ Repetition Penalty

- Frequency penalty: discouraging repeating words too much
- Presence penalty: encourage using different words

# Diversity / Repetition Controlling

[Overview](#)[Documentation](#)[API reference](#)[Examples](#)[Playground](#)[Fine-tuning](#)[Upgrade](#)[Forum](#)[Help](#)

National Taiwan University

## Playground

Your presets

Save

View code

Share



Chat

SYSTEM

You are a helpful assistant.

USER

Enter a user message here.

⊕ Add message

Submit



Model

gpt-3.5-turbo

Temperature

1

Maximum length

256

Stop sequences

Enter sequence and press Tab

Top P

1

Frequency penalty

0

Presence penalty

0



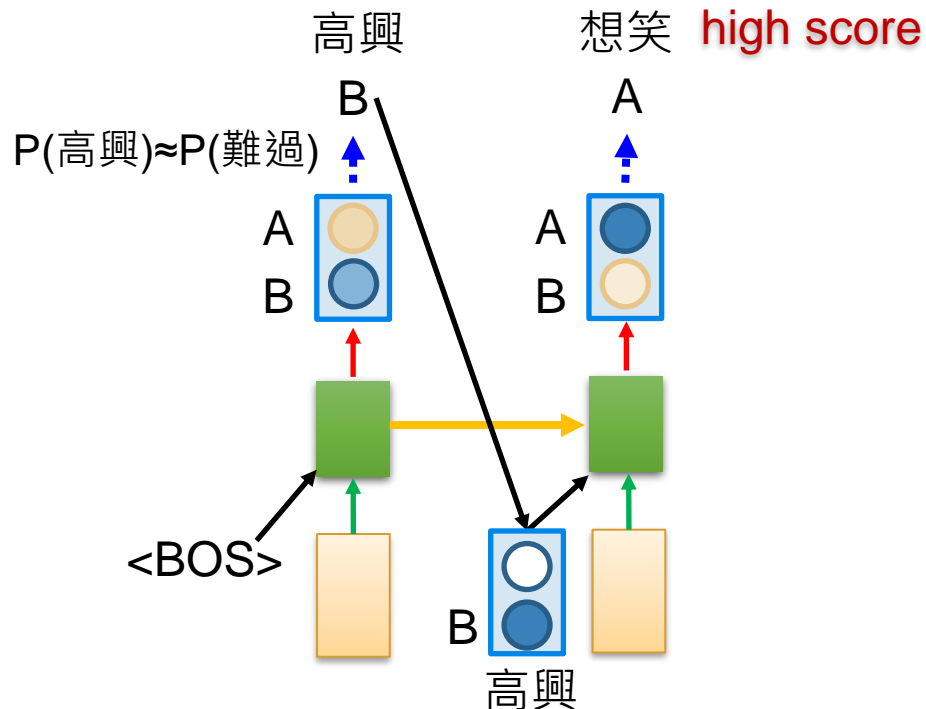
API and Playground requests will not be used to train our models. [Learn more](#)

# Distribution Input

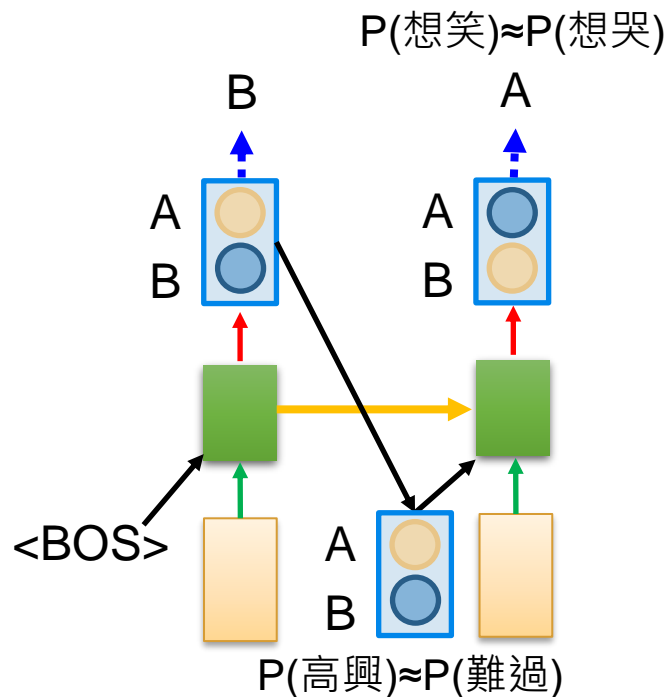
U: 你覺得如何?

M: 高興想笑 or 難過想哭

## One-Hot Input



## Distribution Input



Distribution input may not be good for NLG



# Concluding Remarks

---

- NLG / Conditional NLG

- Decoding Algorithm

- Greedy
- Beam Search
- Sampling
- Top- $k$  Sampling
- Nucleus Sampling

- Generation Controlling

- Temperature
- Frequency Penalty
- Presence Penalty