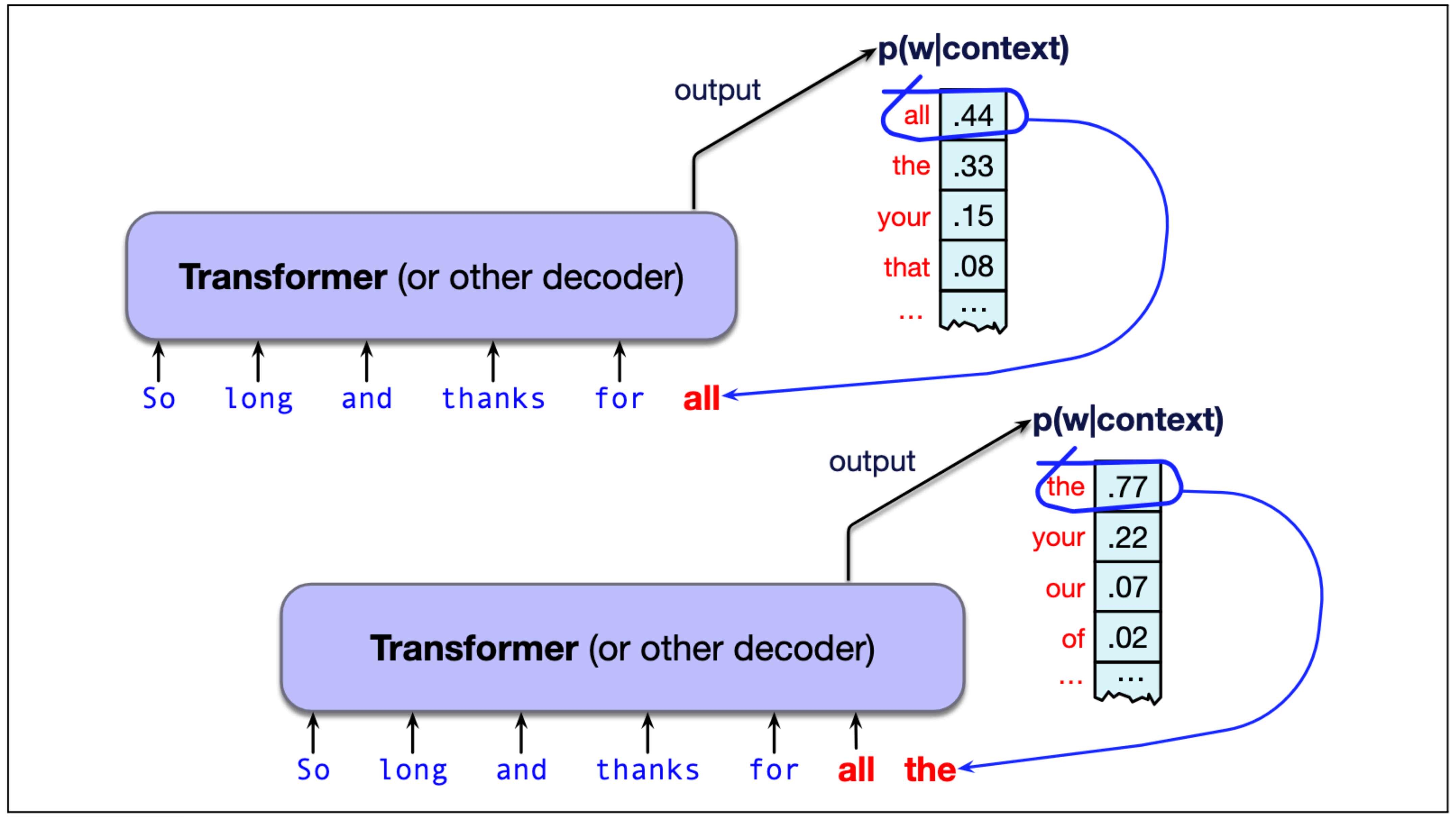
Sampling and Generation

Ling 282/482: Deep Learning for Computational Linguistics

C.M. Downey

Fall 2025

Generation / Decoding

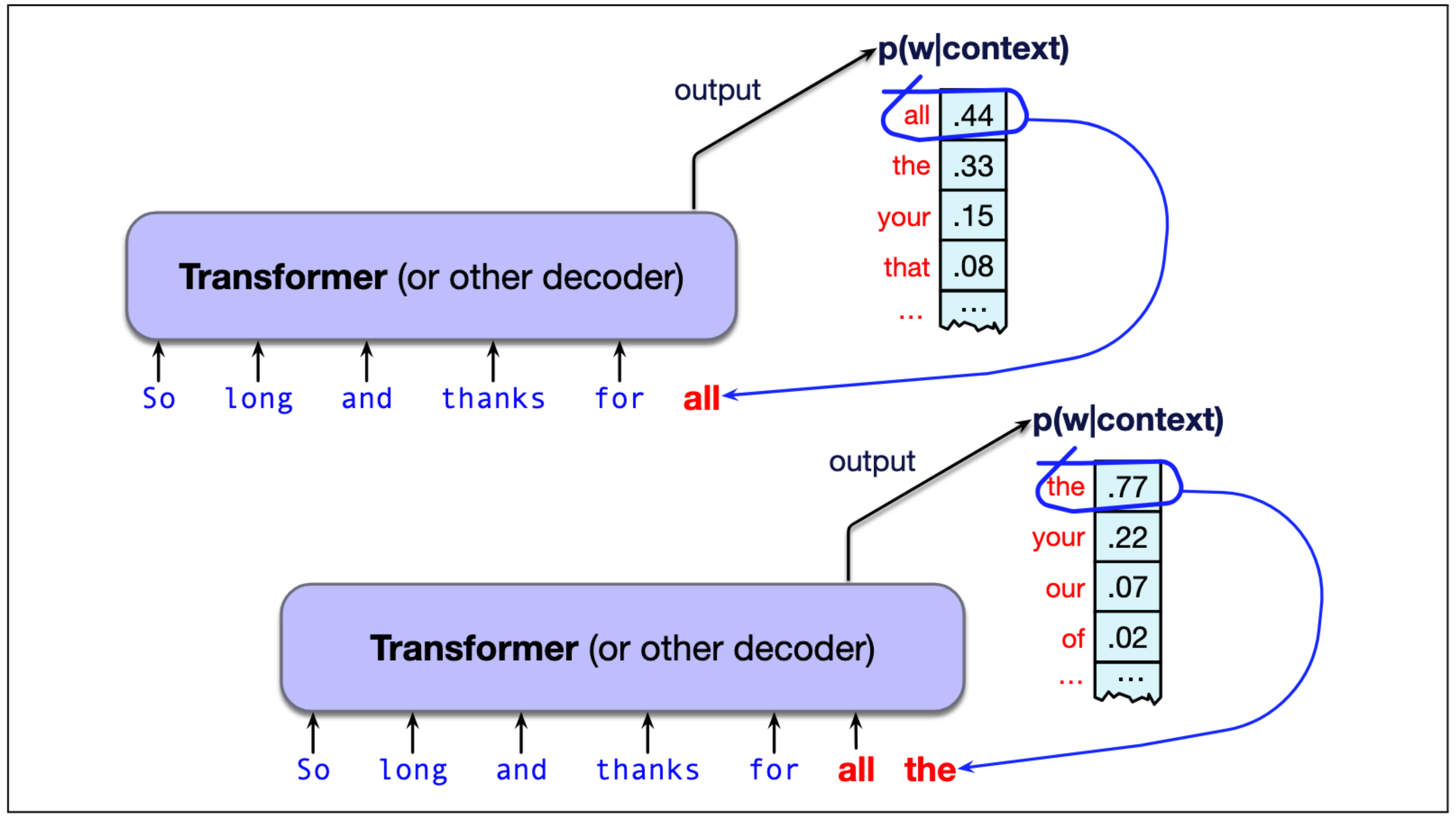


p(w|context)
output
all .44
the .33
your .15
that .08
... ...
Transformer (or other decoder)
So long and thanks for all
p(w|context)
output
the .77
your .22
our .07
of .02
... ...
Transformer (or other decoder)
So long and thanks for all the

Figure 7.2 Turning a predictive model that gives a probability distribution over next words into a generative model by repeatedly sampling from the distribution. The result is a left-to-right (also called autoregressive) language models. As each token is generated, it gets added onto the context as a prefix for generating the next token.

Generation / Decoding

- A Language Model outputs a **probability distribution** over possible words
  - The LM encodes the probability for **all possible sequences**
  - Given this, how do we decide what the **predicted sequence** should be? (This process is often called **"decoding"**)



p(w|context)
output
all .44
the .33
your .15
that .08
… …
Transformer (or other decoder)
So long and thanks for all
p(w|context)
output
the .77
your .22
our .07
of .02
… …
Transformer (or other decoder)
So long and thanks for all the

Figure 7.2 Turning a predictive model that gives a probability distribution over next words into a generative model by repeatedly sampling from the distribution. The result is a left-to-right (also called autoregressive) language models. As each token is generated, it gets added onto the context as a prefix for generating the next token.

Generation / Decoding

- A Language Model outputs a **probability distribution** over possible words
  - The LM encodes the probability for **all possible sequences**
  - Given this, how do we decide what the **predicted sequence** should be? (This process is often called **"decoding"**)
- How do we generate **new sequences?**
  - During training, we always know what the next word should be
  - For many real-world tasks (e.g. Chatbots), we want the model to **generate novel text**

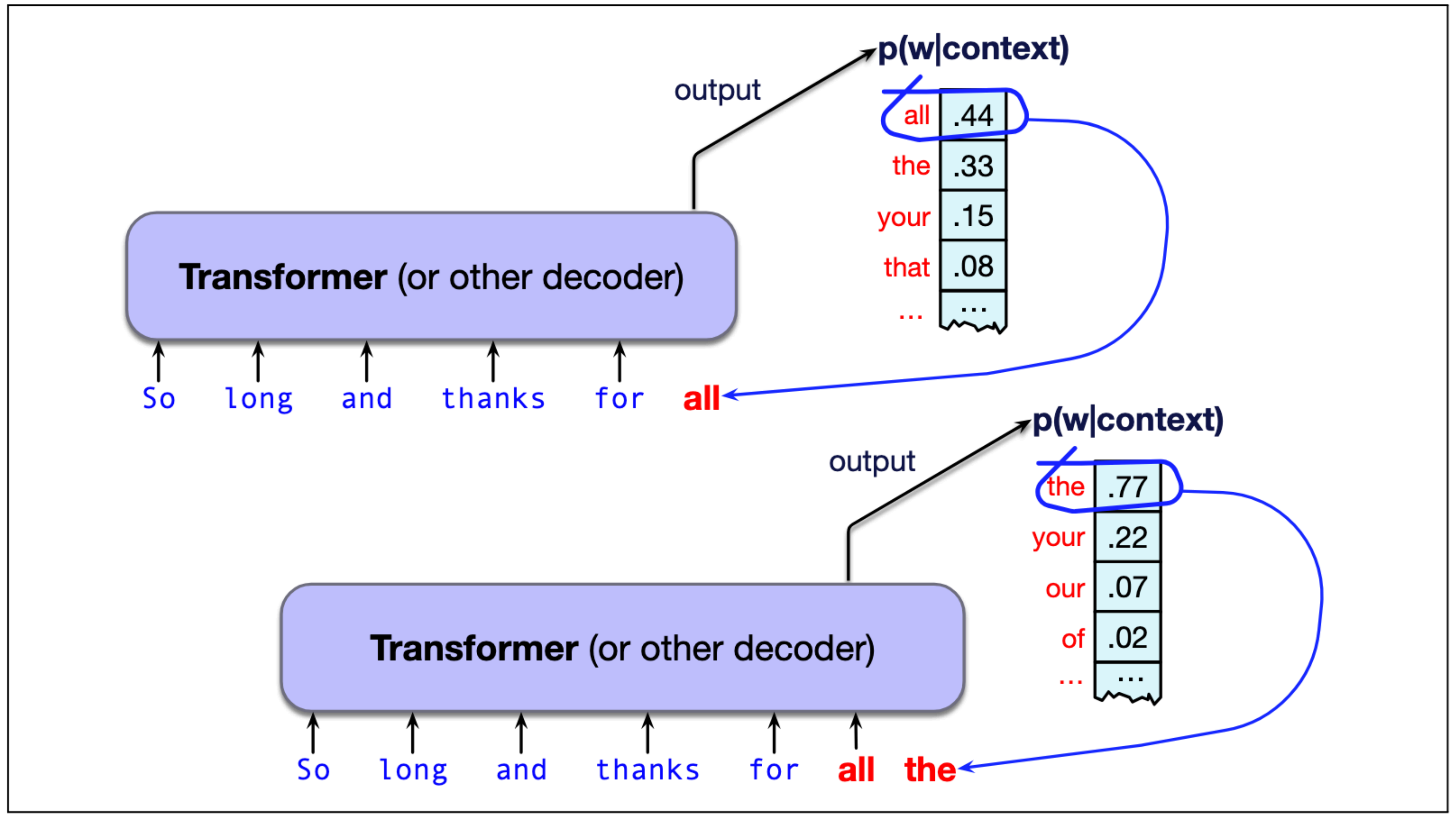

output
p(w|context)
all .44
the .33
your .15
that .08
... ...
Transformer (or other decoder)
So long and thanks for all
output
p(w|context)
the .77
your .22
our .07
of .02
... ...
Transformer (or other decoder)
So long and thanks for all the

Figure 7.2 Turning a predictive model that gives a probability distribution over next words into a generative model by repeatedly sampling from the distribution. The result is a left-to-right (also called autoregressive) language models. As each token is generated, it gets added onto the context as a prefix for generating the next token.

Greedy Decoding

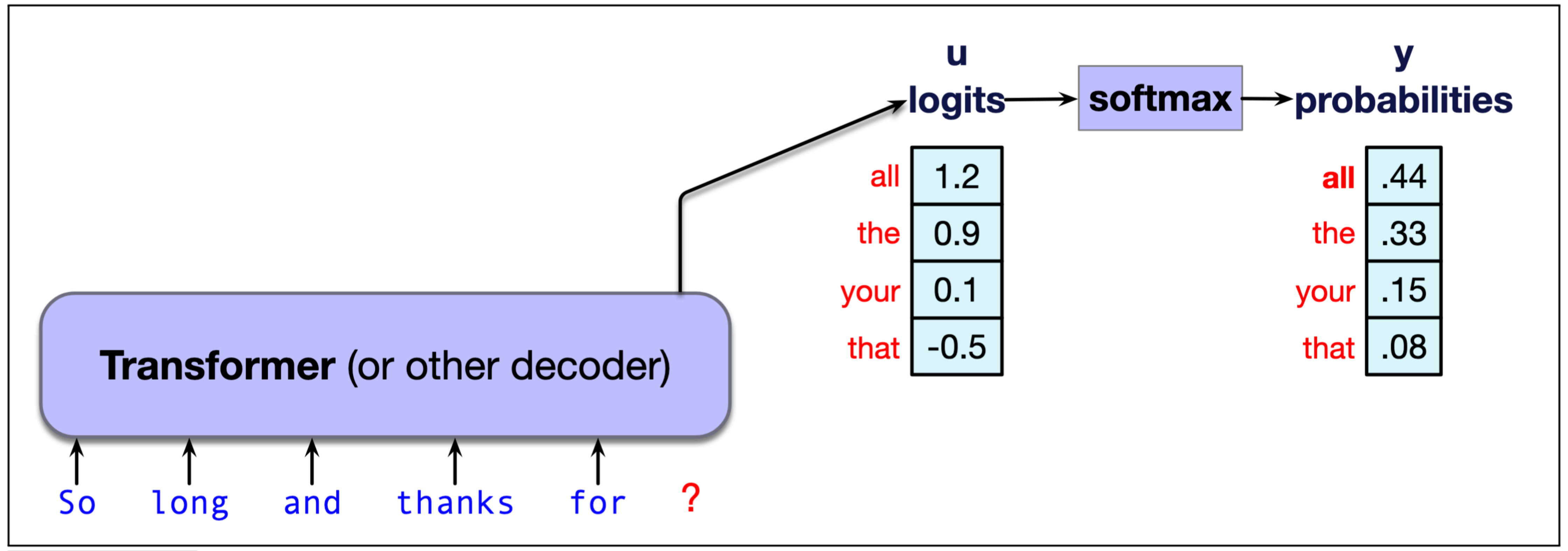


Figure 7.7 Taking the logit vector **u** and using the softmax to create a probability vector **y**.

$$\hat{w}_t = \text{argmax}_{w \in V} P(w \mid w_{<t})$$

Greedy Decoding

- **"Greedy" decoding** is the simplest strategy
  - At each time step, choose the **highest-probability word**
  - "Greedy" because it does **NOT** guarantee the highest-probability sequence

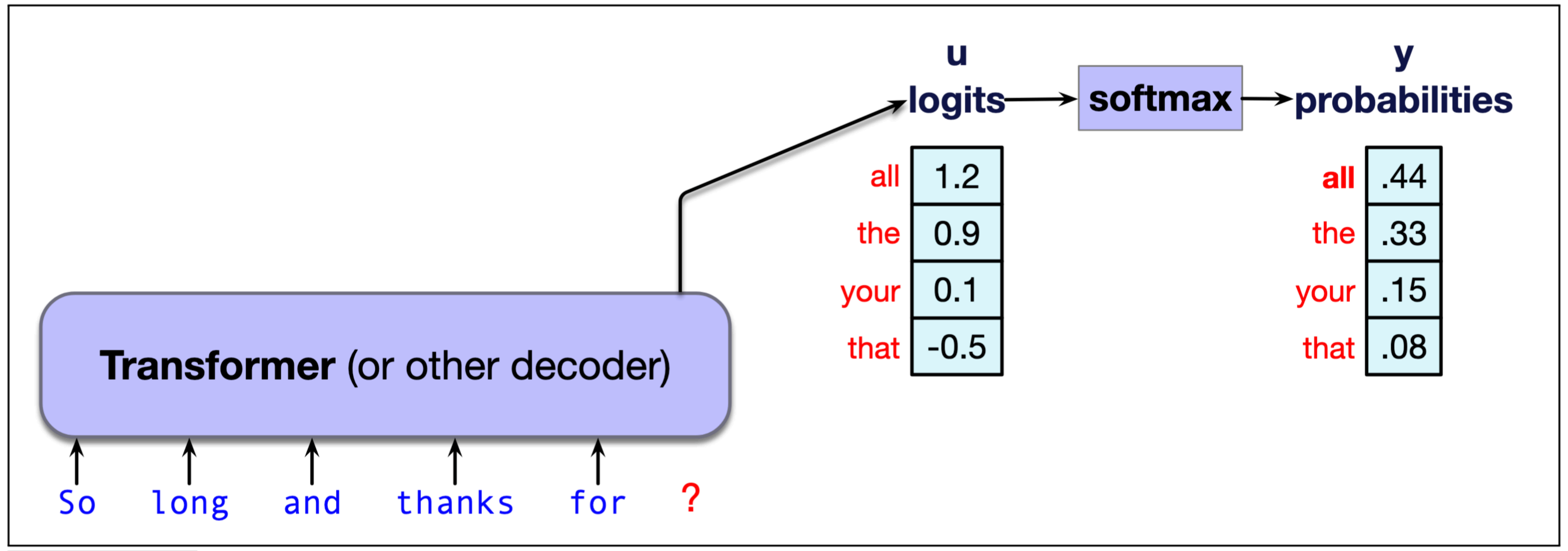


Figure 7.7 Taking the logit vector **u** and using the softmax to create a probability vector **y**.

$$\hat{w}_t = \text{argmax}_{w \in V} P(w \mid w_{<t})$$

Greedy Decoding

- **"Greedy" decoding** is the simplest strategy
  - At each time step, choose the **highest-probability word**
  - "Greedy" because it does **NOT** guarantee the highest-probability sequence
- **No randomness** involved (same context gives the same completion)

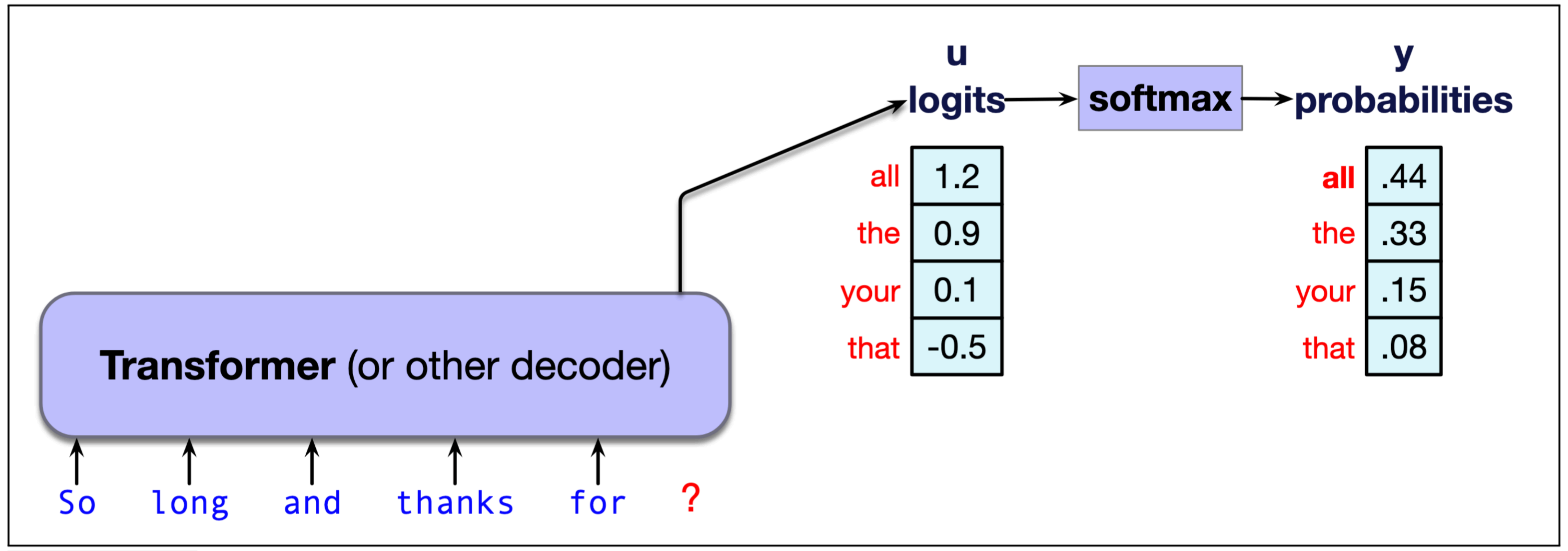


Figure 7.7 Taking the logit vector **u** and using the softmax to create a probability vector **y**.

$$\hat{w}_t = \text{argmax}_{w \in V} P(w \mid w_{<t})$$

Greedy Decoding

- **"Greedy" decoding** is the simplest strategy
  - At each time step, choose the **highest-probability word**
  - "Greedy" because it does **NOT** guarantee the highest-probability sequence
- **No randomness** involved (same context gives the same completion)
- Tends to generate **boring or repetitive** text
  - Or text that's been **exactly copied** from the training data

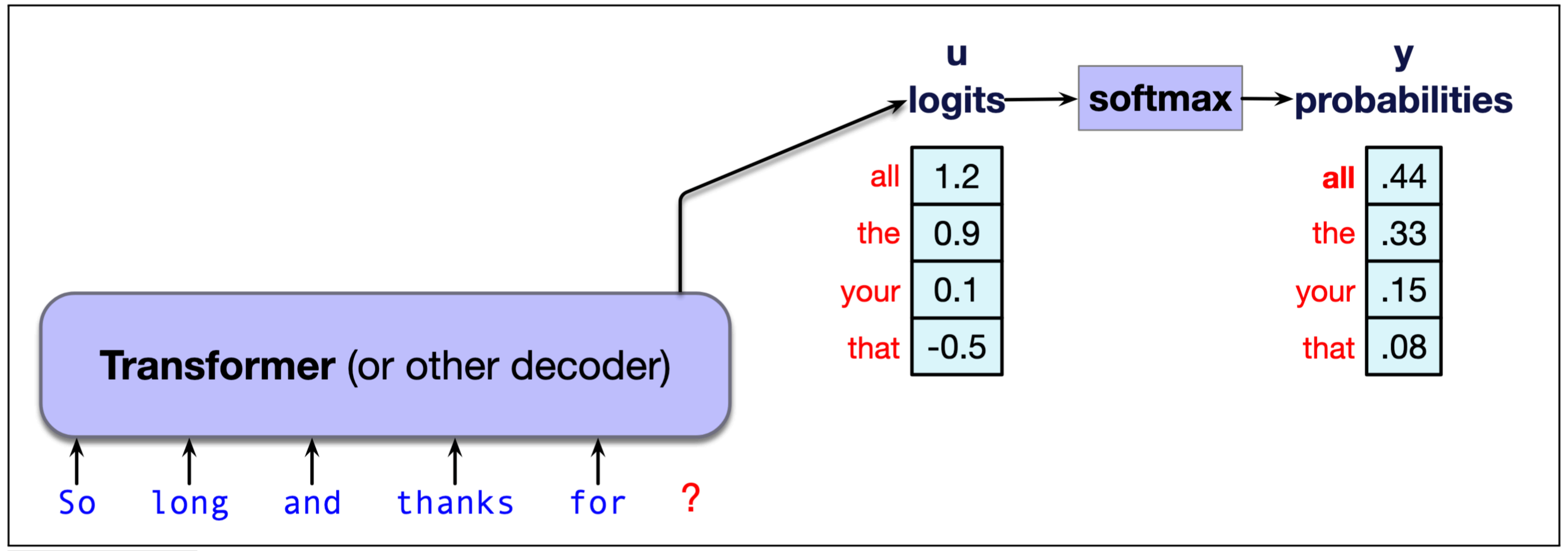


Figure 7.7 Taking the logit vector **u** and using the softmax to create a probability vector **y**.

$$\hat{w}_t = \text{argmax}_{w \in V} P(w \mid w_{<t})$$

UNIVERSITY of ROCHESTER

Random Sampling

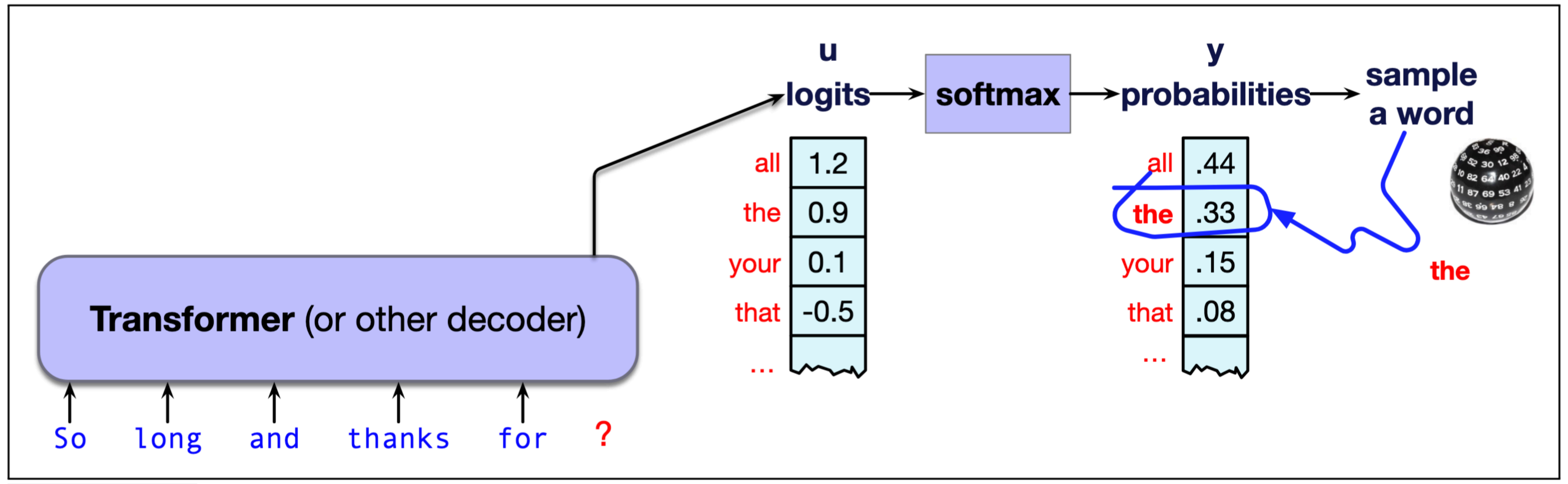


Figure 7.9 Random multinomial sampling: we randomly chose a word according to its probability.

```
$i \leftarrow 1$
$w_i \sim p(w)$
**while** $w_i$ != EOS
    $i \leftarrow i + 1$
    $w_i \sim p(w_i \mid w_{<i})$
```

Random Sampling

- **Sampling:** taking **random draws** from a **probability distribution**

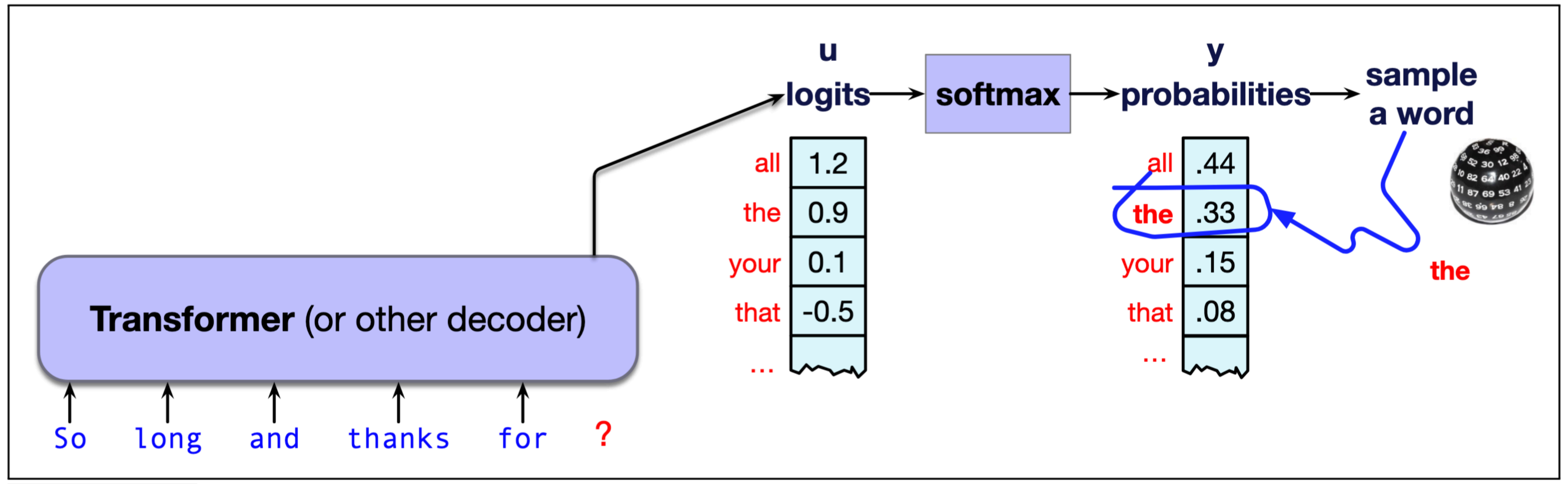


Figure 7.9 Random multinomial sampling: we randomly chose a word according to its probability.

```
i←1
$w_i \sim$ p($w$)
**while** $w_i$ != EOS
    i←i + 1
    $w_i \sim$ p($w_i \mid w_{<i}$)
```

Random Sampling

- **Sampling:** taking **random draws** from a **probability distribution**
  - For LMs: sample from **distribution over possible words**

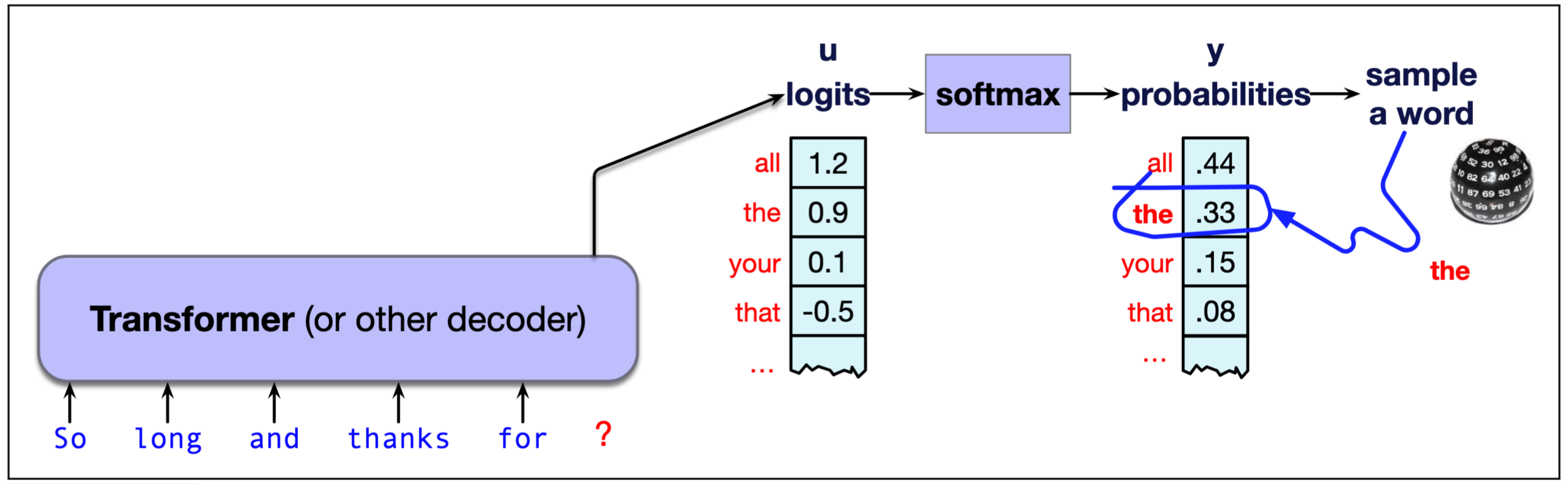


Figure 7.9 Random multinomial sampling: we randomly chose a word according to its probability.

```
i←1
$w_i \sim p(w)$
while $w_i$ != EOS
    i←i + 1
    $w_i \sim p(w_i \mid w_{<i})$
```

Random Sampling

- **Sampling:** taking **random draws** from a **probability distribution**
  - For LMs: sample from **distribution over possible words**
  - Stop when the **End-Of-Sequence** token is reached, or define a **maximum sequence length**

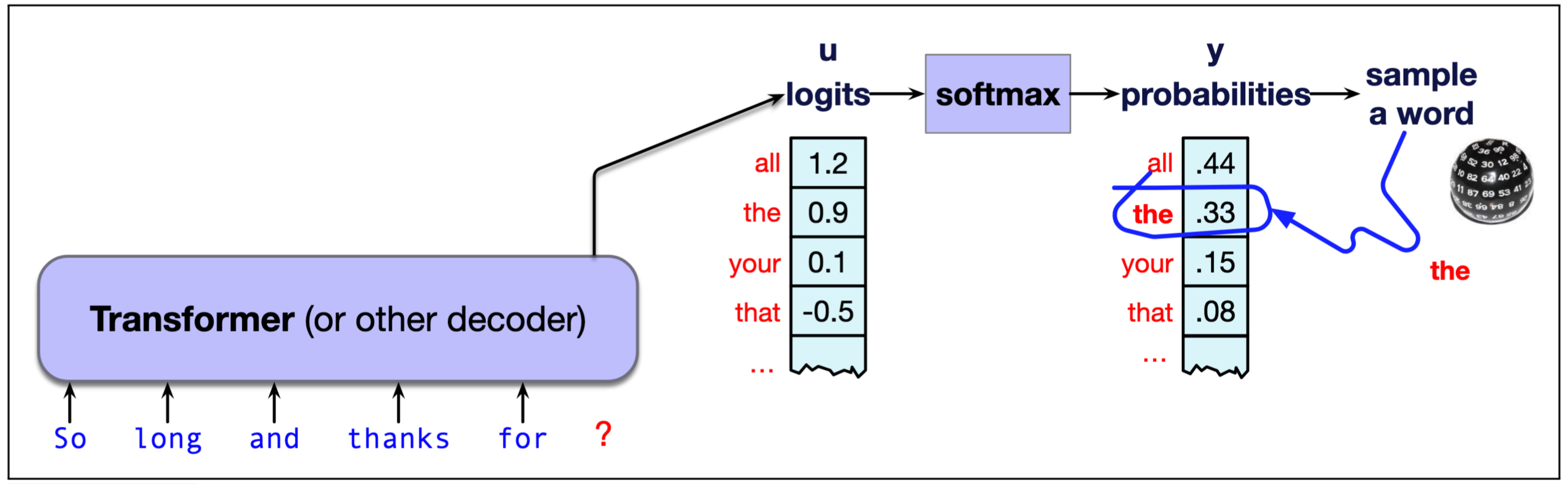


Figure 7.9 Random multinomial sampling: we randomly chose a word according to its probability.

```
$\mathrm{i} \leftarrow 1$
$w_i \sim \mathrm{p}(w)$
while $w_i$ != EOS
    $\mathrm{i} \leftarrow \mathrm{i} + 1$
    $w_i \sim \mathrm{p}(w_i \mid w_{<i})$
```

Random Sampling

- **Sampling:** taking **random draws** from a **probability distribution**
  - For LMs: sample from **distribution over possible words**
  - Stop when the **End-Of-Sequence** token is reached, or define a **maximum sequence length**
- Each word has the **probability assigned by the LM** of being generated

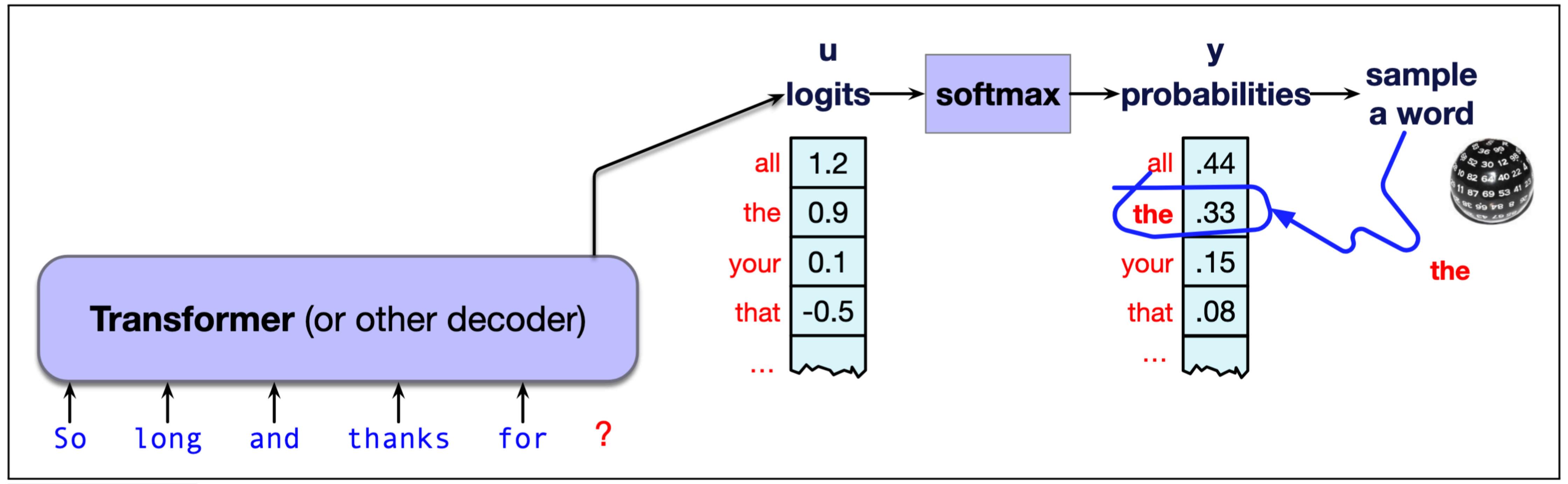


Figure 7.9 Random multinomial sampling: we randomly chose a word according to its probability.

```
$i \leftarrow 1$
$w_i \sim p(w)$
while $w_i$ != EOS
    $i \leftarrow i + 1$
    $w_i \sim p(w_i \mid w_{<i})$
```

Random Sampling

- **Sampling:** taking **random draws** from a **probability distribution**
    - For LMs: sample from **distribution over possible words**
    - Stop when the **End-Of-Sequence** token is reached, or define a **maximum sequence length**
- Each word has the **probability assigned by the LM** of being generated
    - Pros: relatively **"interesting"**, novel text generation

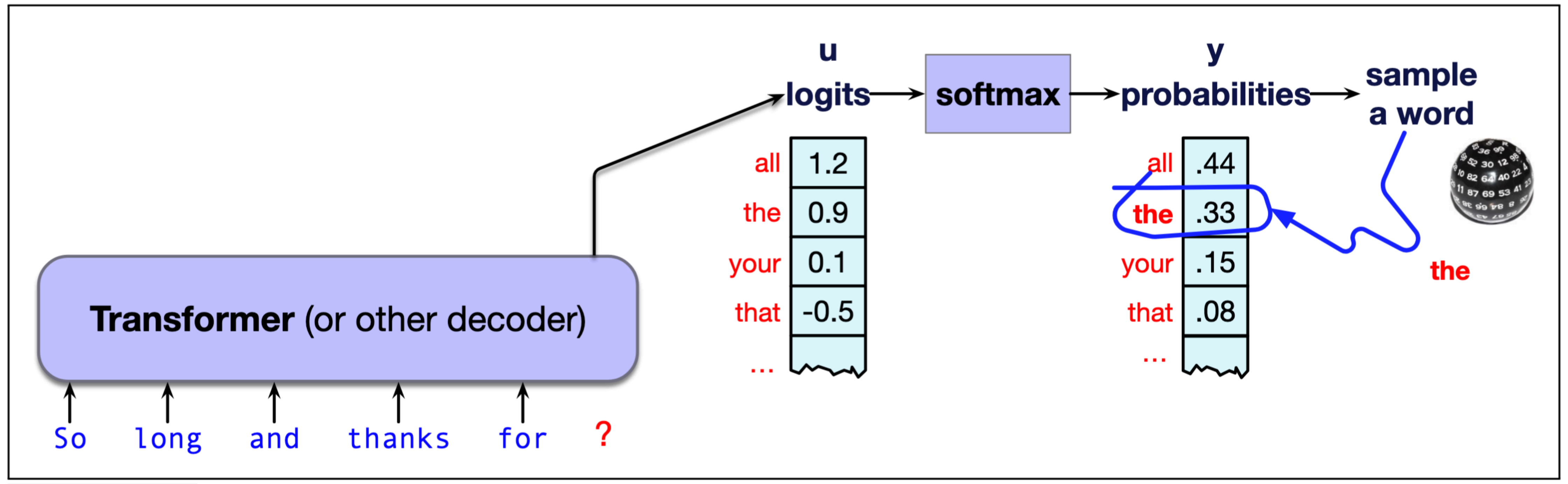


Figure 7.9 Random multinomial sampling: we randomly chose a word according to its probability.

```
$i \leftarrow 1$
$w_i \sim p(w)$
while $w_i$ != EOS
    $i \leftarrow i + 1$
    $w_i \sim p(w_i \mid w_{<i})$
```

Random Sampling

- **Sampling:** taking **random draws** from a **probability distribution**
    - For LMs: sample from **distribution over possible words**
    - Stop when the **End-Of-Sequence** token is reached, or define a **maximum sequence length**
- Each word has the **probability assigned by the LM** of being generated
    - Pros: relatively **"interesting"**, novel text generation
    - Cons: high chance of **generating nonsense** (because of the **long tail** of **low-probability choices**)

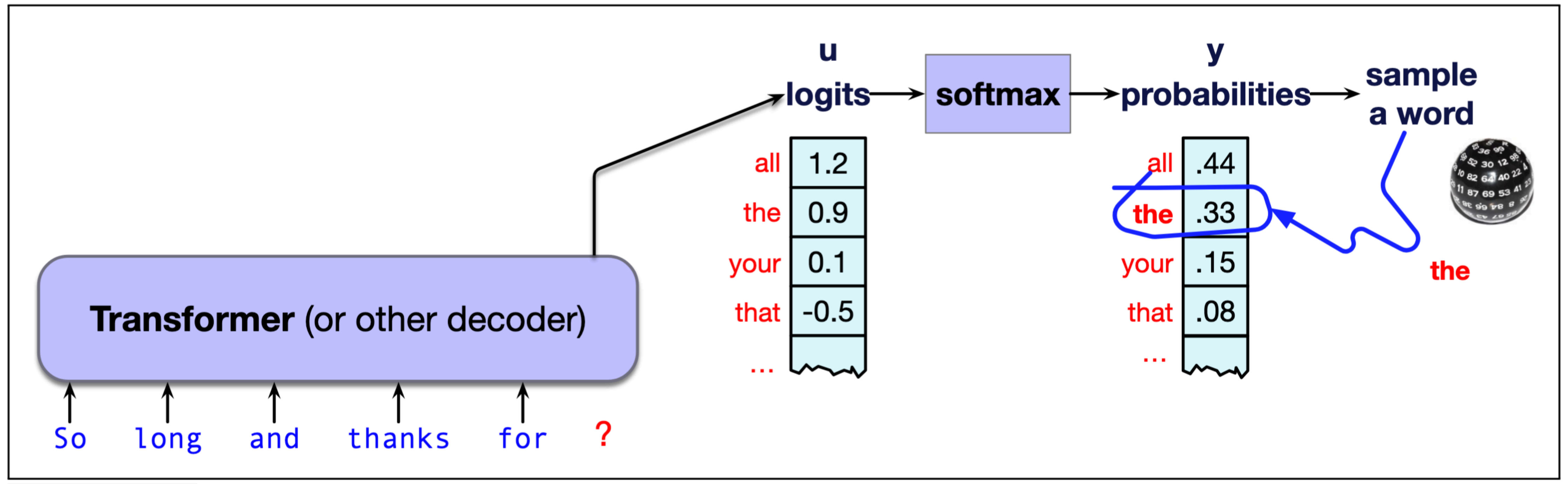


Figure 7.9 Random multinomial sampling: we randomly chose a word according to its probability.

```
$i \leftarrow 1$
$w_i \sim p(w)$
while $w_i$ != EOS
    $i \leftarrow i + 1$
    $w_i \sim p(w_i \mid w_{<i})$
```

Random Sampling

- **Sampling:** taking **random draws** from a **probability distribution**
  - For LMs: sample from **distribution over possible words**
  - Stop when the **End-Of-Sequence** token is reached, or define a **maximum sequence length**
- Each word has the **probability assigned by the LM** of being generated
  - Pros: relatively **"interesting"**, novel text generation
  - Cons: high chance of **generating nonsense** (because of the **long tail** of **low-probability choices**)
- Is there something in-between this and greedy?

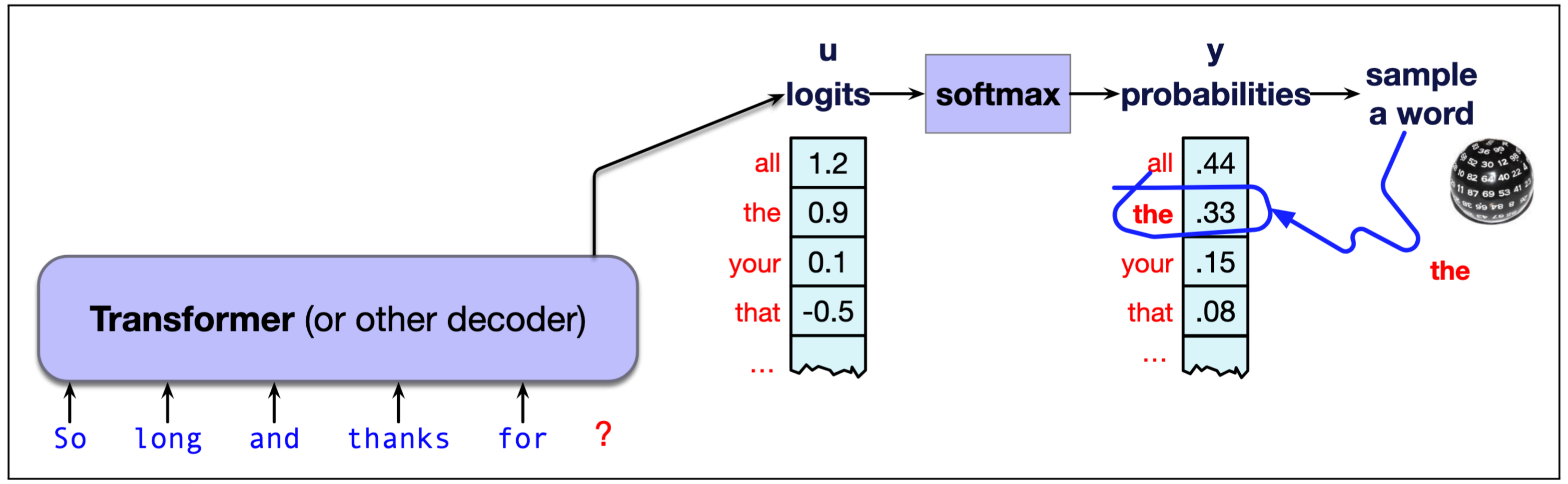
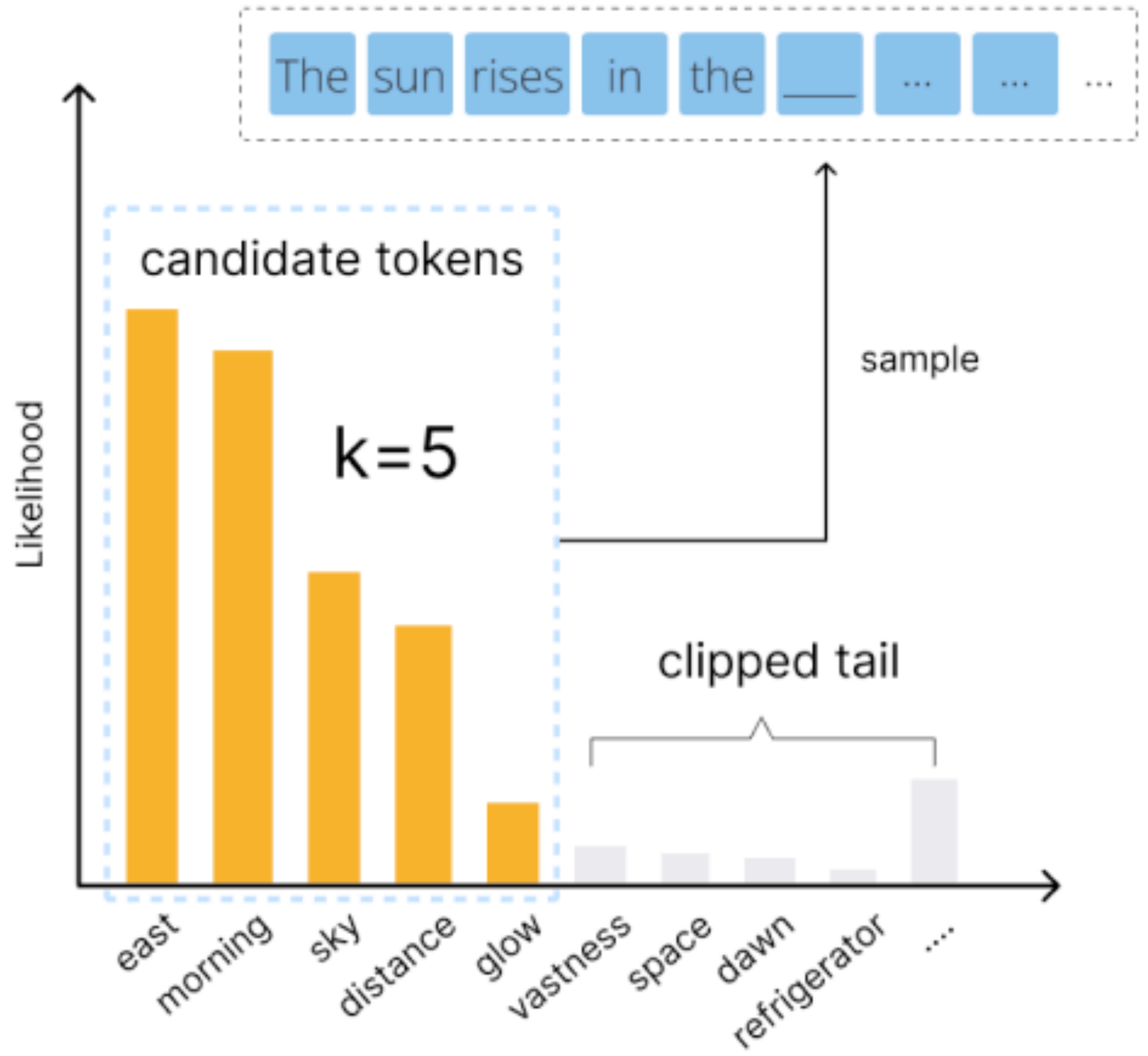


Figure 7.9 Random multinomial sampling: we randomly chose a word according to its probability.

```
$i \leftarrow 1$
$w_i \sim\ p(w)$
while $w_i$ != EOS
    $i \leftarrow i + 1$
    $w_i \sim\ p(w_i \mid w_{<i})$
```

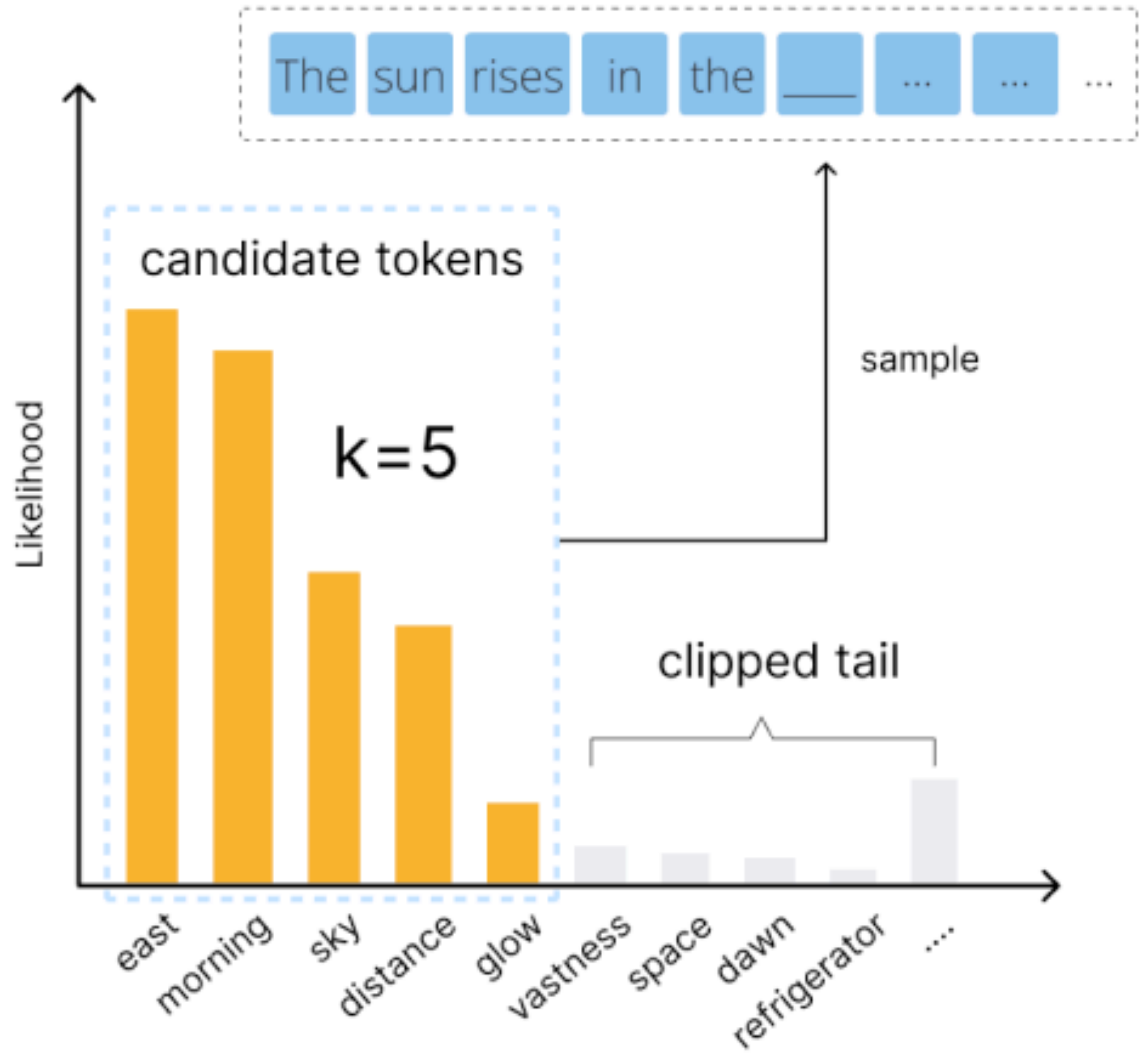
Top-k Sampling



source

Top-k Sampling

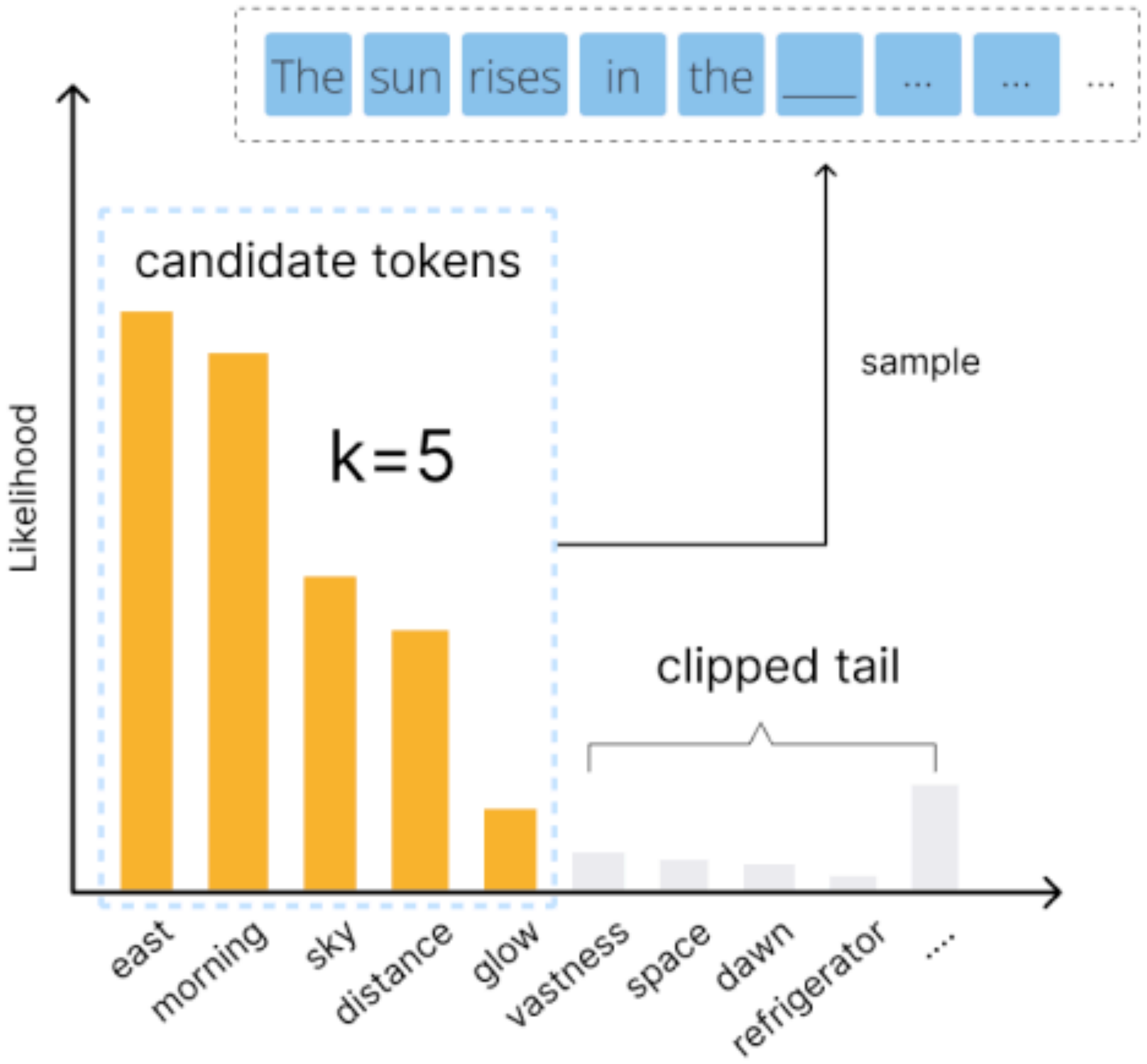
- Instead of considering the whole distribution, what about the **top few words?**



source

Top-k Sampling

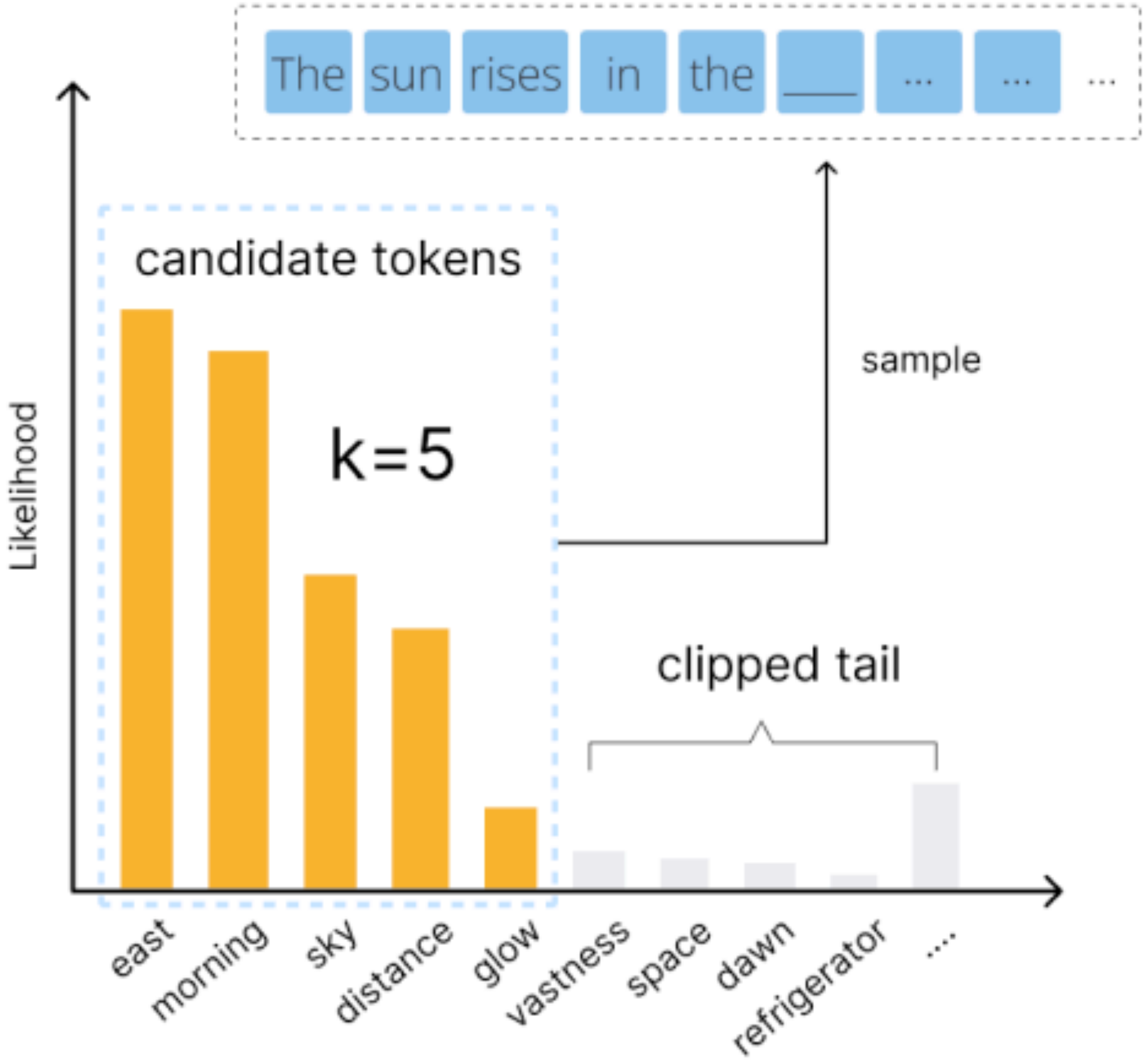
- Instead of considering the whole distribution, what about the **top few words?**
- Top-k Sampling:
  - Take the ***k*** **highest-probability** words
  - Sample **among these words** according to their probability



source

Top-k Sampling

- Instead of considering the whole distribution, what about the **top few words?**
- Top-k Sampling:
  - Take the ***k* highest-probability** words
  - Sample **among these words** according to their probability
- **Cuts off** the long tail of the distribution



source

Top-p Sampling

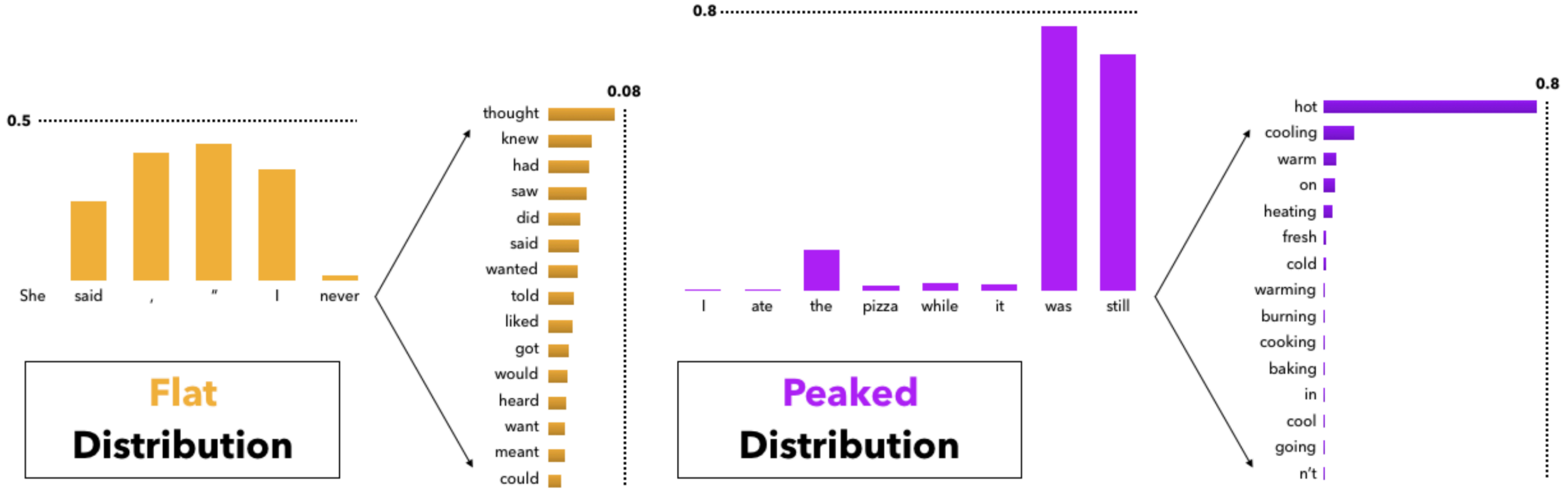


Figure 5: The probability mass assigned to partial human sentences. Flat distributions lead to many moderately probable tokens, while peaked distributions concentrate most probability mass into just a few tokens. The presence of flat distributions makes the use of a small k in top-k sampling problematic, while the presence of peaked distributions makes large k's problematic.

[Holtzman et al (2020)]

Top-p Sampling

- Problem with top-k: probability distributions can look **very different**

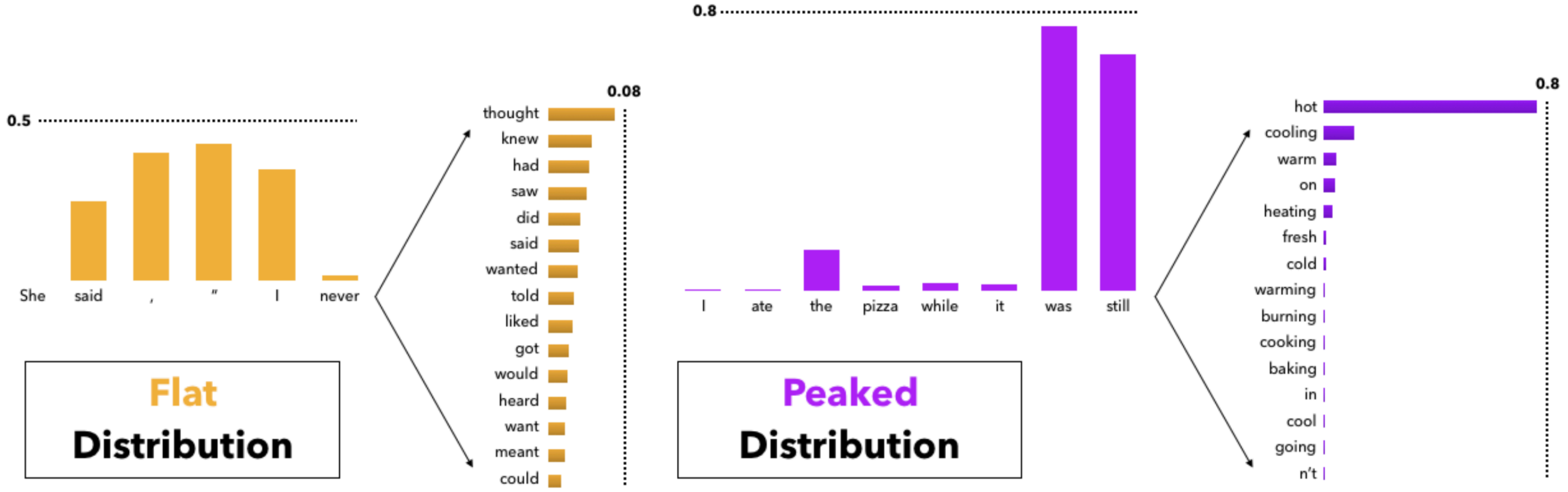


Figure 5: The probability mass assigned to partial human sentences. Flat distributions lead to many moderately probable tokens, while peaked distributions concentrate most probability mass into just a few tokens. The presence of flat distributions makes the use of a small k in top-k sampling problematic, while the presence of peaked distributions makes large k's problematic.

Holtzman et al (2020)

Top-p Sampling

- Problem with top-k: probability distributions can look **very different**
  - Sometimes the top k will make up the **majority** of the **probability mass** (peaked)

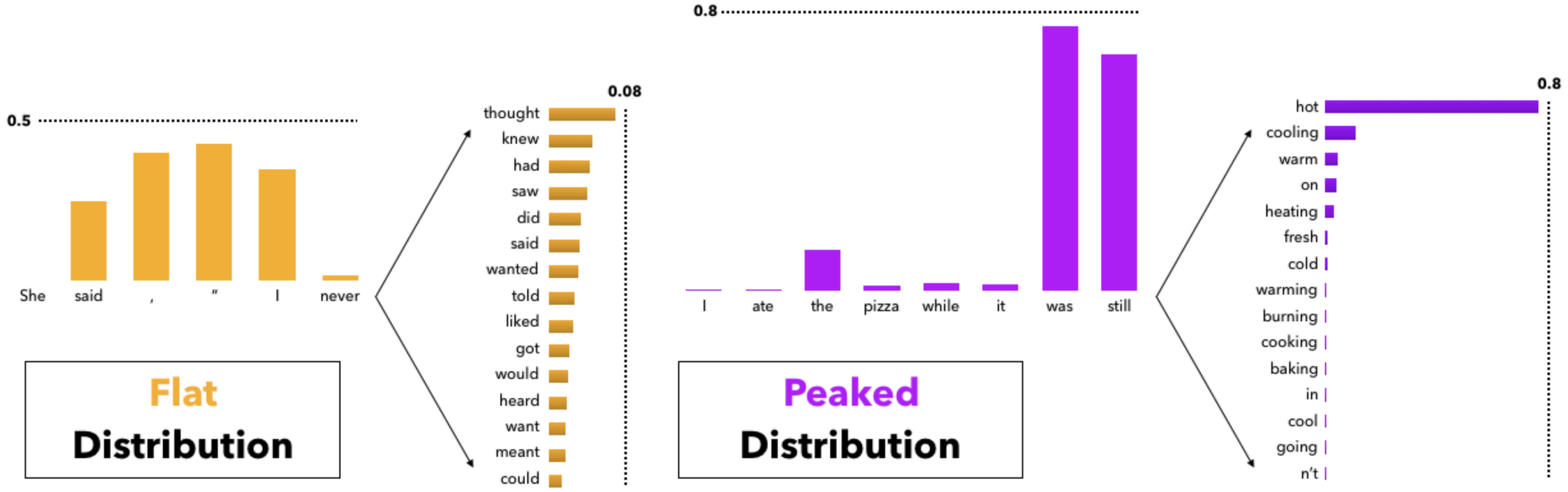


Figure 5: The probability mass assigned to partial human sentences. Flat distributions lead to many moderately probable tokens, while peaked distributions concentrate most probability mass into just a few tokens. The presence of flat distributions makes the use of a small k in top-k sampling problematic, while the presence of peaked distributions makes large k's problematic.

Holtzman et al (2020)

Top-p Sampling

- Problem with top-k: probability distributions can look **very different**
  - Sometimes the top k will make up the **majority** of the **probability mass** (peaked)
  - Other times the distribution is **spread out** (flat)

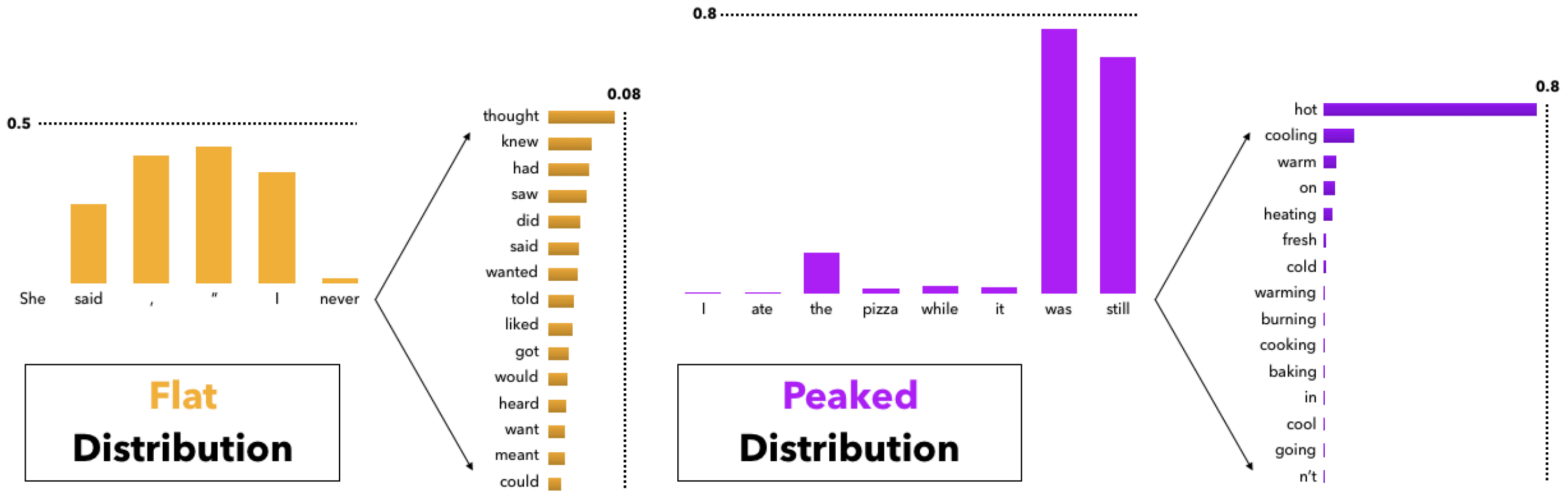


Figure 5: The probability mass assigned to partial human sentences. Flat distributions lead to many moderately probable tokens, while peaked distributions concentrate most probability mass into just a few tokens. The presence of flat distributions makes the use of a small k in top-k sampling problematic, while the presence of peaked distributions makes large k's problematic.

Holtzman et al (2020)

Top-p Sampling

- Problem with top-k: probability distributions can look **very different**
  - Sometimes the top k will make up the **majority** of the **probability mass** (peaked)
  - Other times the distribution is **spread out** (flat)
  - (Hard to find a *k* that always works well)

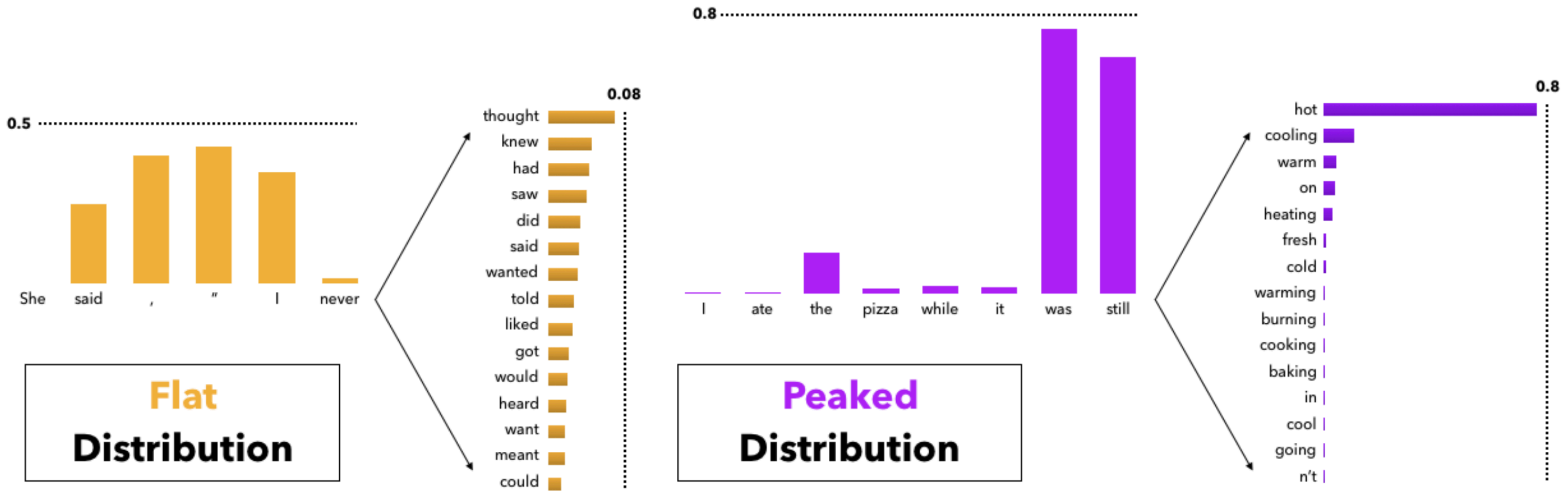


Figure 5: The probability mass assigned to partial human sentences. Flat distributions lead to many moderately probable tokens, while peaked distributions concentrate most probability mass into just a few tokens. The presence of flat distributions makes the use of a small k in top-k sampling problematic, while the presence of peaked distributions makes large k's problematic.

Holtzman et al (2020)

Top-p Sampling

- Problem with top-k: probability distributions can look **very different**
  - Sometimes the top k will make up the **majority** of the **probability mass** (peaked)
  - Other times the distribution is **spread out** (flat)
  - (Hard to find a *k* that always works well)
- Top-p (AKA **nucleus**) **sampling:** truncate the distribution to the **top probability mass** (e.g. 0.2 / 20%)

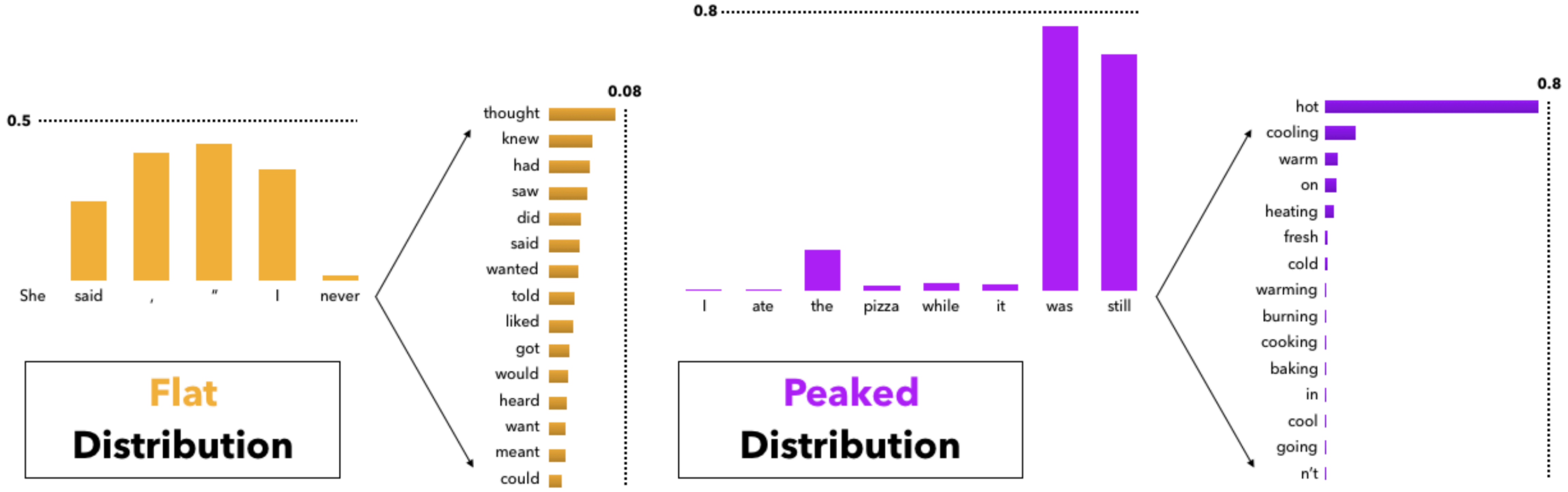


Figure 5: The probability mass assigned to partial human sentences. Flat distributions lead to many moderately probable tokens, while peaked distributions concentrate most probability mass into just a few tokens. The presence of flat distributions makes the use of a small k in top-k sampling problematic, while the presence of peaked distributions makes large k's problematic.

Holtzman et al (2020)

Top-p Sampling

- Problem with top-k: probability distributions can look **very different**
  - Sometimes the top k will make up the **majority** of the **probability mass** (peaked)
  - Other times the distribution is **spread out** (flat)
  - (Hard to find a *k* that always works well)
- Top-p (AKA **nucleus**) **sampling:** truncate the distribution to the **top probability mass** (e.g. 0.2 / 20%)
  - Adaptable to different distribution shapes

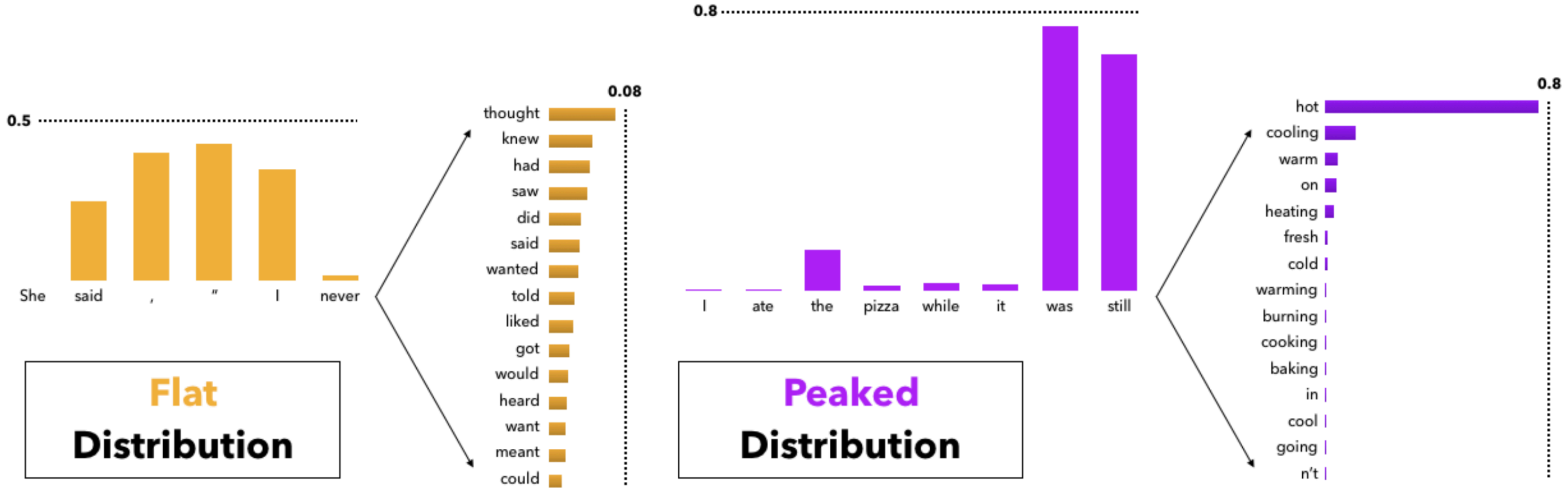
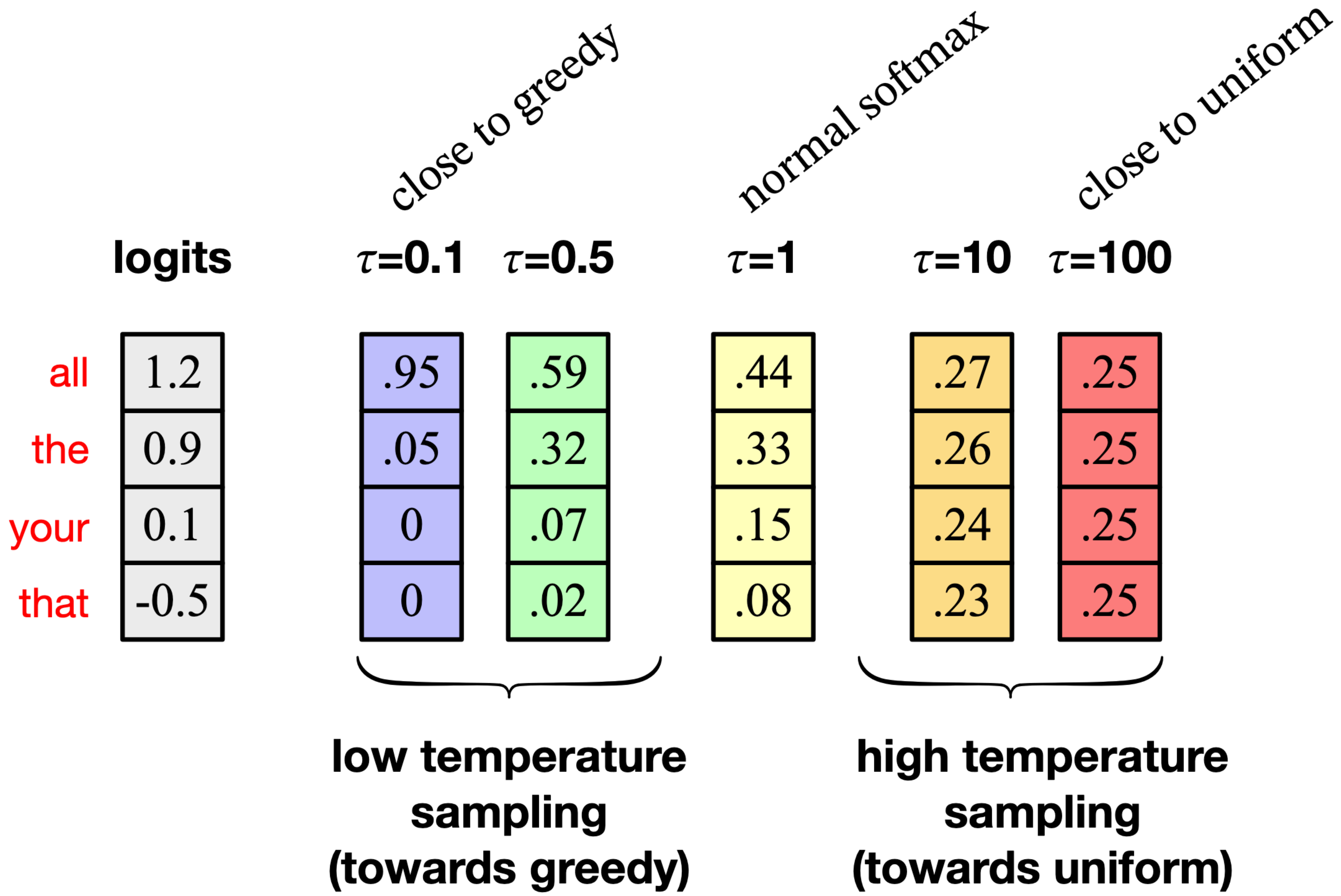


Figure 5: The probability mass assigned to partial human sentences. Flat distributions lead to many moderately probable tokens, while peaked distributions concentrate most probability mass into just a few tokens. The presence of flat distributions makes the use of a small k in top-k sampling problematic, while the presence of peaked distributions makes large k's problematic.

Holtzman et al (2020)

Softmax Temperature

$$\text{probs} = \text{softmax}(\mathbf{x}/\tau)$$



close to greedy

normal softmax

close to uniform

	logits	**τ=0.1**	**τ=0.5**	**τ=1**	**τ=10**	**τ=100**
all	1.2	.95	.59	.44	.27	.25
the	0.9	.05	.32	.33	.26	.25
your	0.1	0	.07	.15	.24	.25
that	-0.5	0	.02	.08	.23	.25

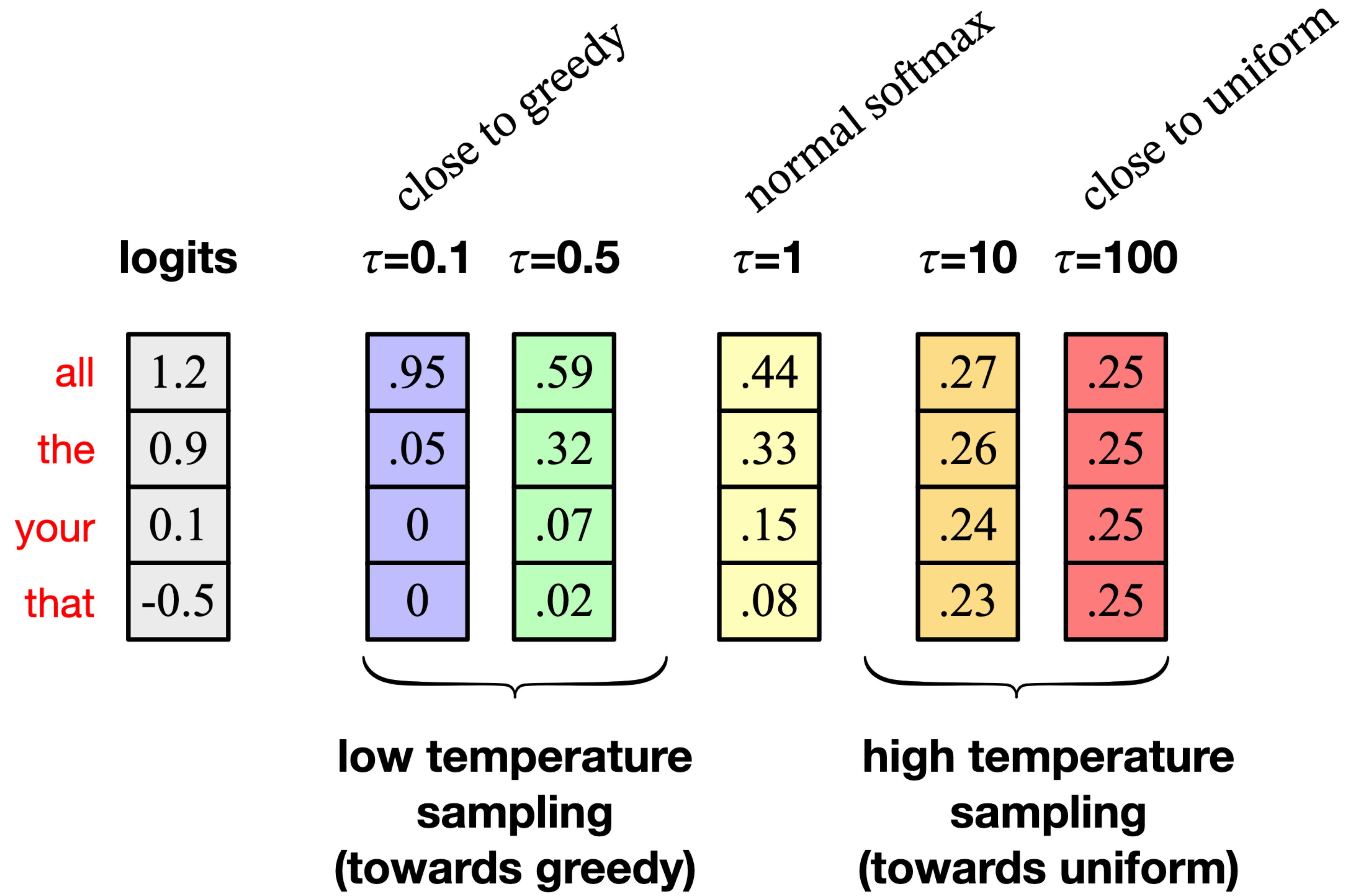
low temperature sampling (towards greedy)

high temperature sampling (towards uniform)

Softmax Temperature

- The "peakiness" of a distribution can be adjusted with parameter called **temperature** (τ)

$$\text{probs} = \text{softmax}(\mathbf{x}/\tau)$$



close to greedy

normal softmax

close to uniform

	logits	**τ=0.1**	**τ=0.5**	**τ=1**	**τ=10**	**τ=100**
all	1.2	.95	.59	.44	.27	.25
the	0.9	.05	.32	.33	.26	.25
your	0.1	0	.07	.15	.24	.25
that	-0.5	0	.02	.08	.23	.25

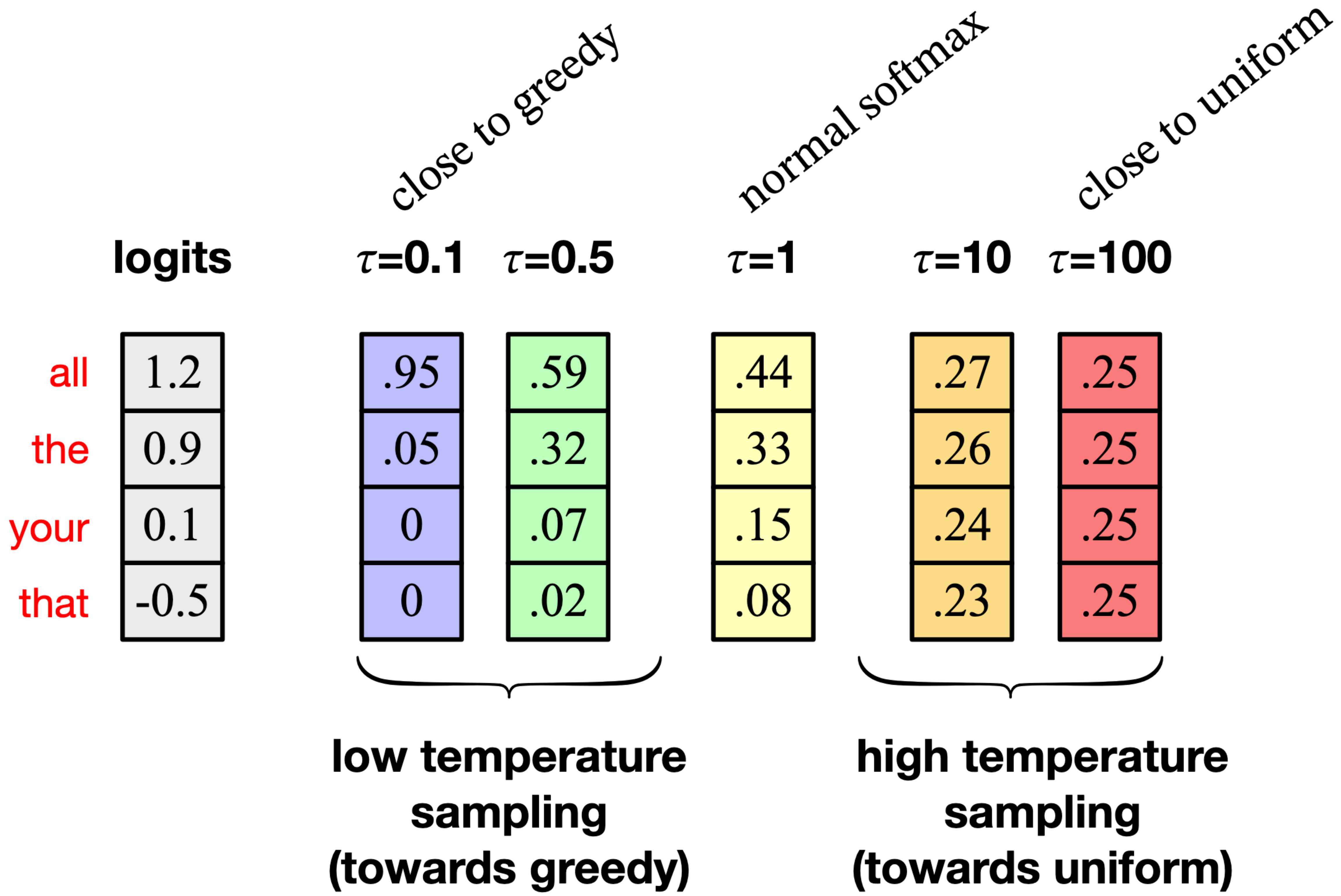
low temperature sampling (towards greedy)

high temperature sampling (towards uniform)

Softmax Temperature

- The "peakiness" of a distribution can be adjusted with parameter called **temperature** (τ)
  - **Low** temperature → **more peaky** / close to greedy sampling

$$\text{probs} = \text{softmax}(\mathbf{x}/\tau)$$



	logits	**τ=0.1**	**τ=0.5**	**τ=1**	**τ=10**	**τ=100**
all	1.2	.95	.59	.44	.27	.25
the	0.9	.05	.32	.33	.26	.25
your	0.1	0	.07	.15	.24	.25
that	-0.5	0	.02	.08	.23	.25

close to greedy
normal softmax
close to uniform

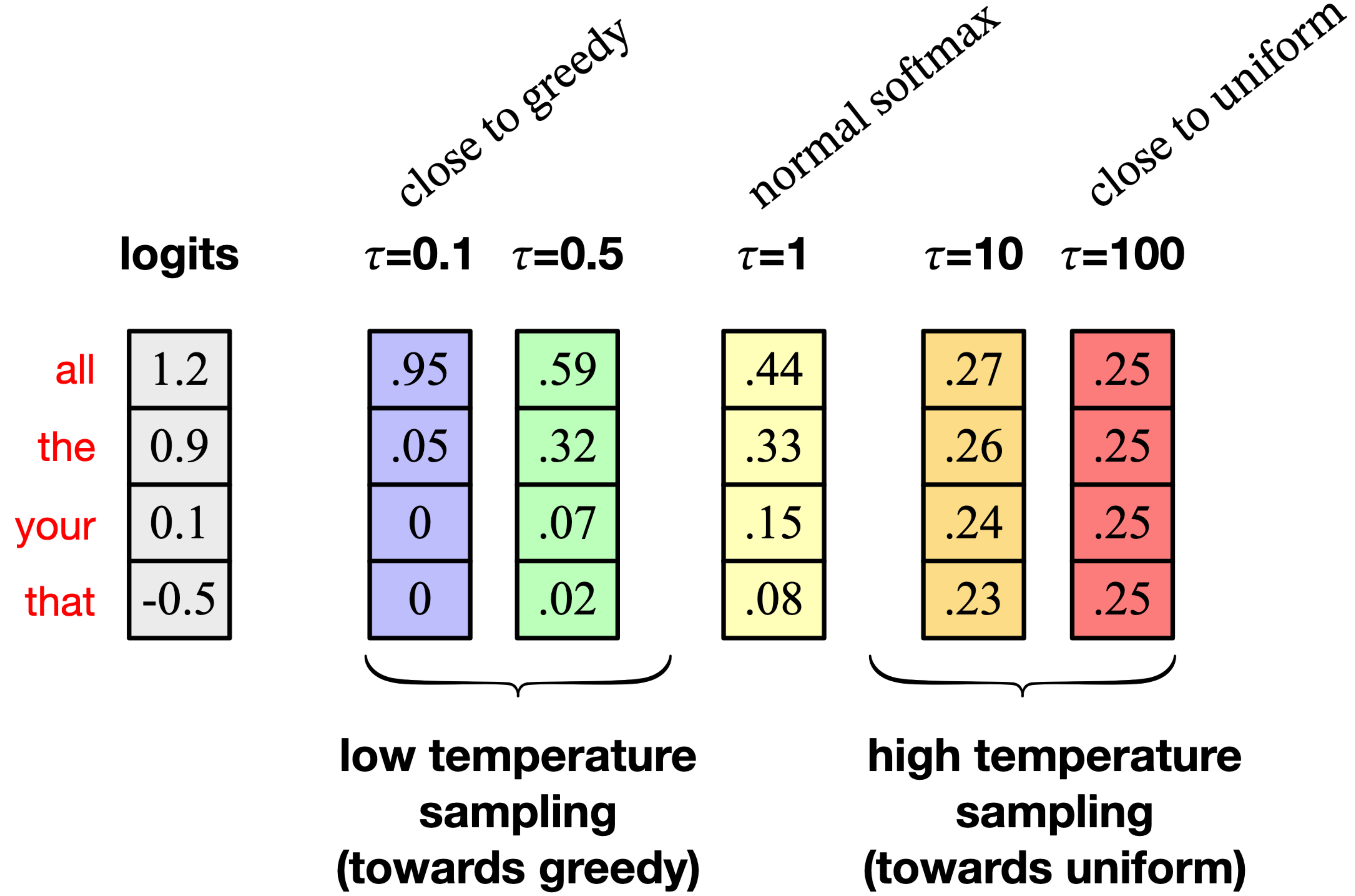
low temperature sampling (towards greedy)

high temperature sampling (towards uniform)

Softmax Temperature

- The "peakiness" of a distribution can be adjusted with parameter called **temperature** (τ)
  - **Low** temperature → **more peaky** / close to greedy sampling
  - **High** temperature → **more flat** / close to uniform distribution

$$\text{probs} = \text{softmax}(\mathbf{x}/\tau)$$



	logits	τ=0.1	τ=0.5	τ=1	τ=10	τ=100
		close to greedy		normal softmax		close to uniform
all	1.2	.95	.59	.44	.27	.25
the	0.9	.05	.32	.33	.26	.25
your	0.1	0	.07	.15	.24	.25
that	-0.5	0	.02	.08	.23	.25

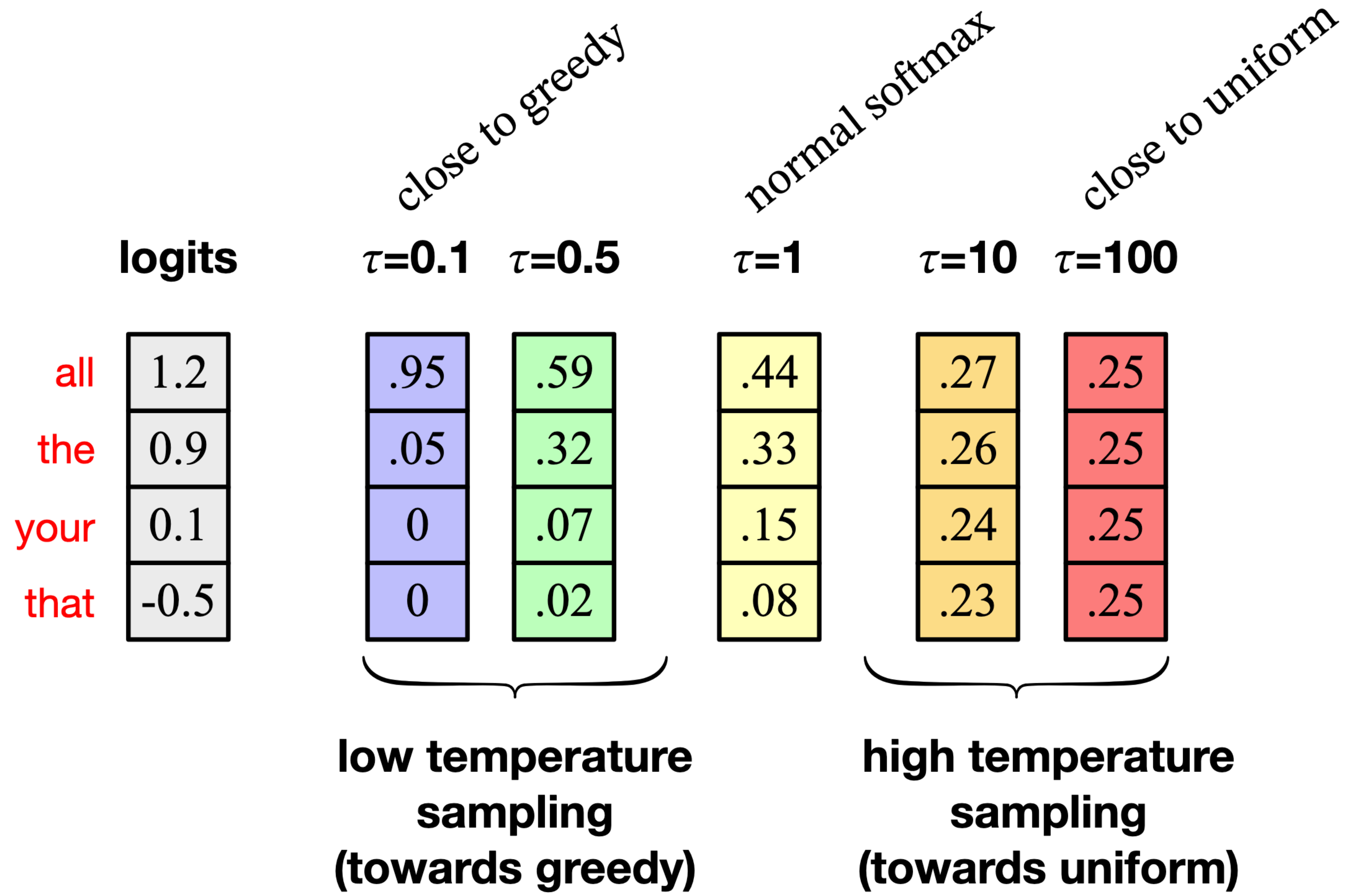
low temperature sampling (towards greedy)

high temperature sampling (towards uniform)

Softmax Temperature

- The "peakiness" of a distribution can be adjusted with parameter called **temperature** (τ)
  - **Low** temperature → **more peaky** / close to greedy sampling
  - **High** temperature → **more flat** / close to uniform distribution
  - $\tau = 1.0$ → **regular softmax**

$$\text{probs} = \text{softmax}(\mathbf{x}/\tau)$$



	logits	**τ=0.1**	**τ=0.5**	**τ=1**	**τ=10**	**τ=100**
all	1.2	.95	.59	.44	.27	.25
the	0.9	.05	.32	.33	.26	.25
your	0.1	0	.07	.15	.24	.25
that	-0.5	0	.02	.08	.23	.25

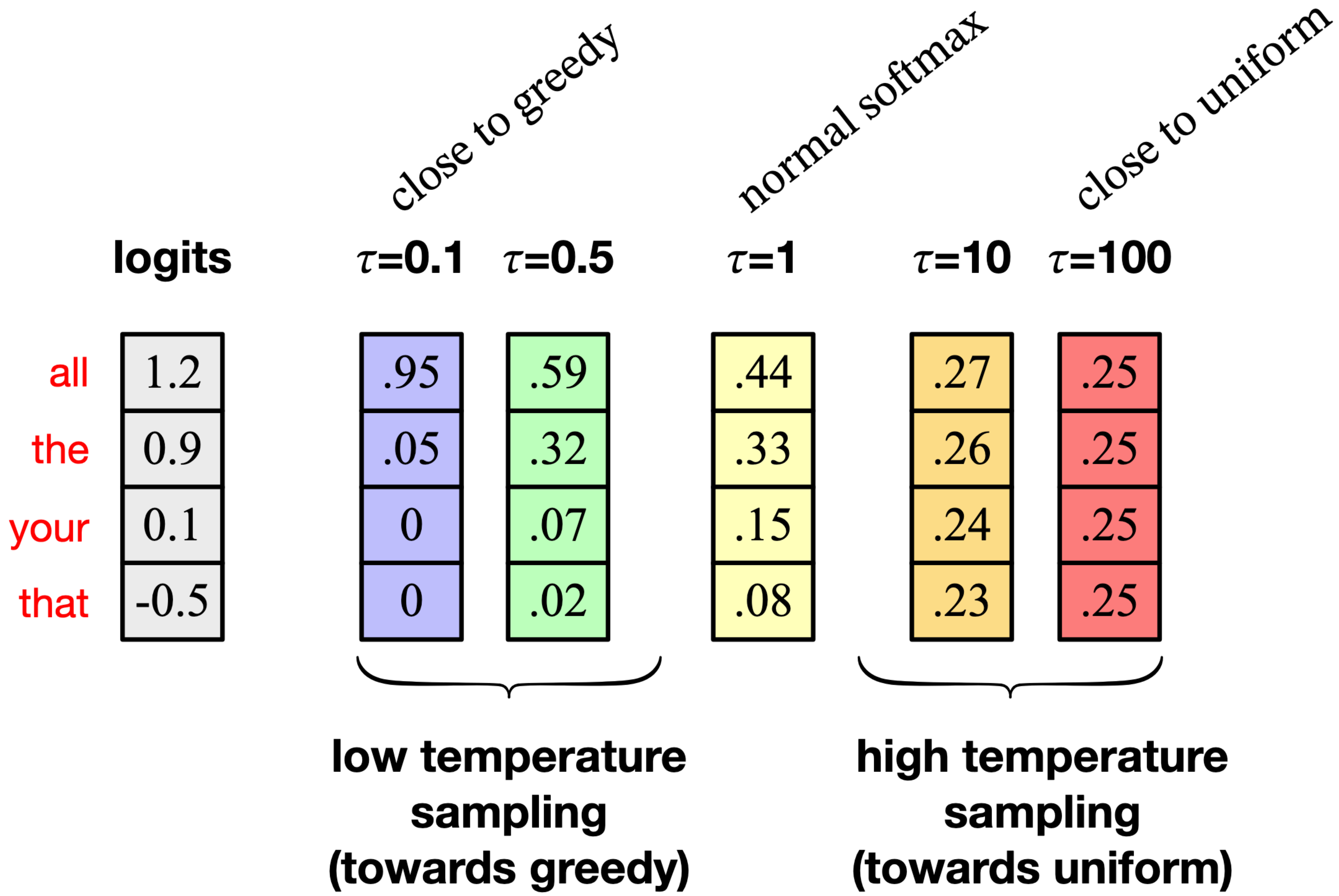
low temperature sampling (towards greedy)

high temperature sampling (towards uniform)

Softmax Temperature

- The "peakiness" of a distribution can be adjusted with parameter called **temperature** (τ)
  - **Low** temperature → **more peaky** / close to greedy sampling
  - **High** temperature → **more flat** / close to uniform distribution
  - $\tau = 1.0$ → **regular softmax**
- Can be **tuned** to give more/less deterministic outputs

$$\text{probs} = \text{softmax}(\mathbf{x}/\tau)$$



close to greedy

normal softmax

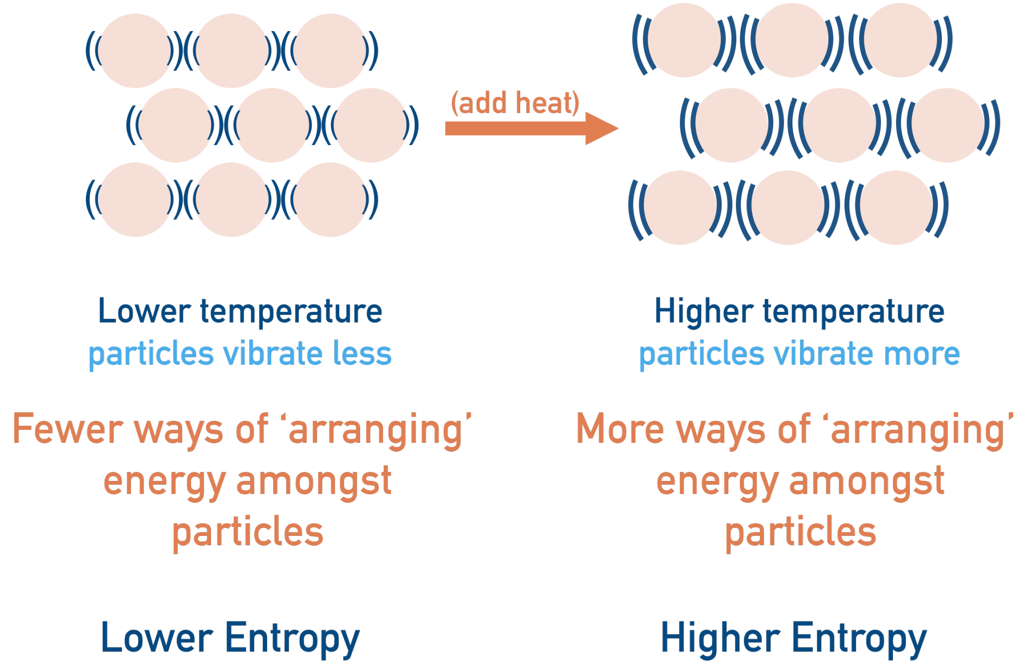
close to uniform

	logits	**τ=0.1**	**τ=0.5**	**τ=1**	**τ=10**	**τ=100**
all	1.2	.95	.59	.44	.27	.25
the	0.9	.05	.32	.33	.26	.25
your	0.1	0	.07	.15	.24	.25
that	-0.5	0	.02	.08	.23	.25

low temperature sampling (towards greedy)

high temperature sampling (towards uniform)

Softmax Temperature



(add heat)

Lower temperature
particles vibrate less

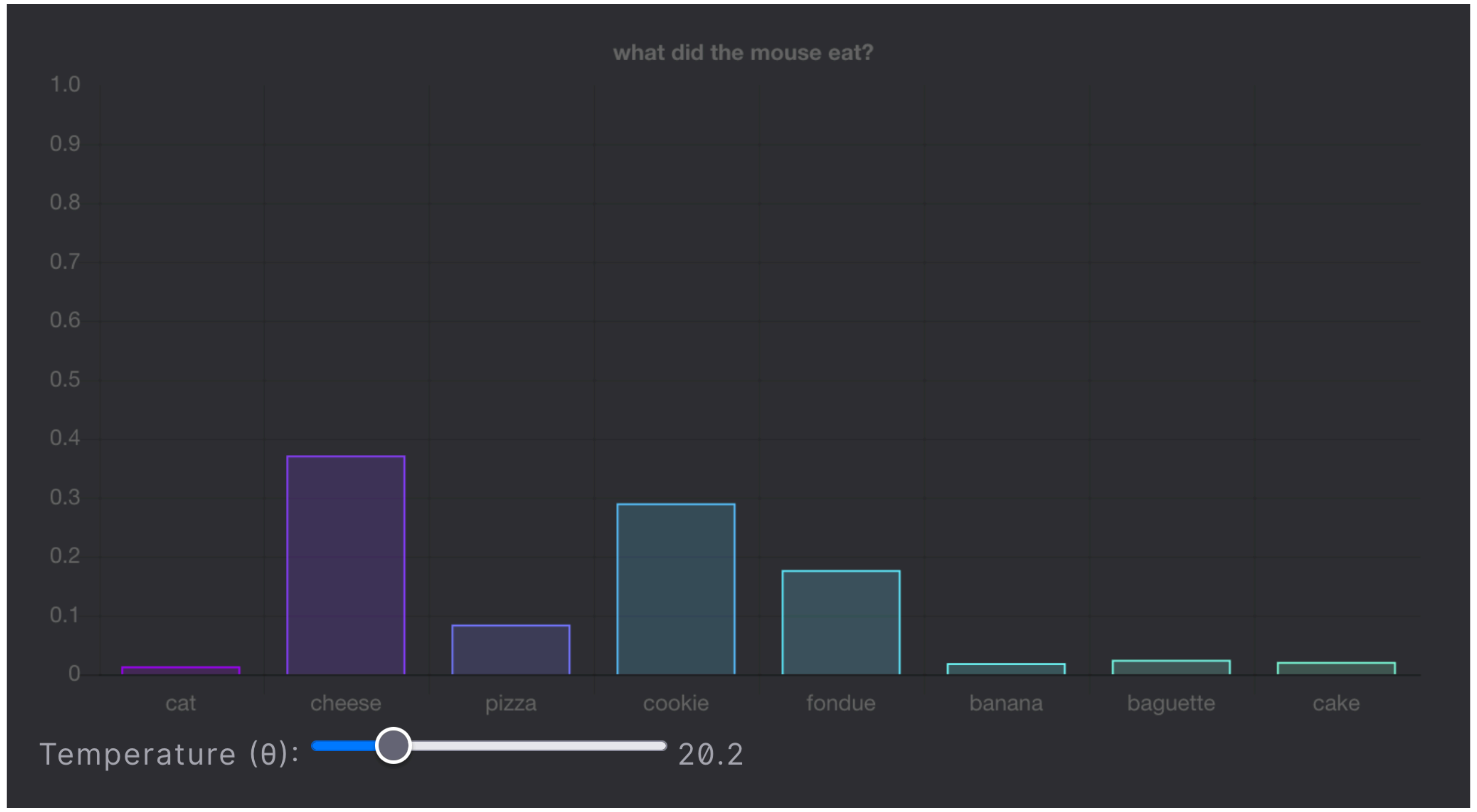
Fewer ways of 'arranging' energy amongst particles

Lower Entropy

Higher temperature
particles vibrate more

More ways of 'arranging' energy amongst particles

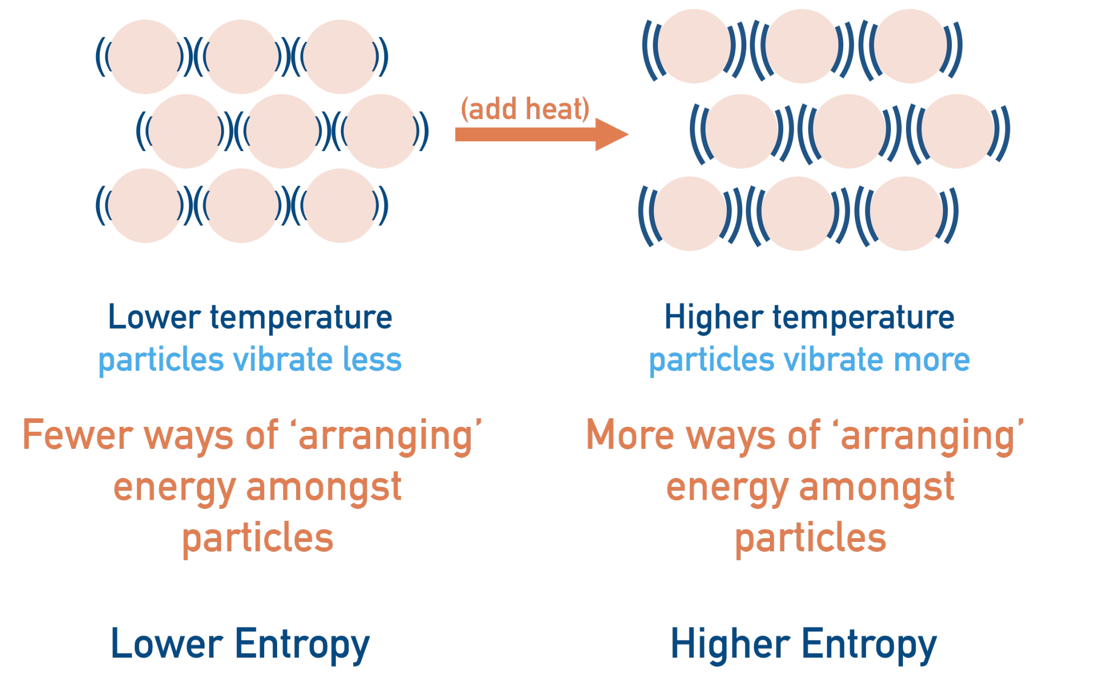
Higher Entropy



Temperature (θ): 20.2

Softmax Temperature

- Takes inspiration from temperature in **physics**



(add heat)

Lower temperature
particles vibrate less

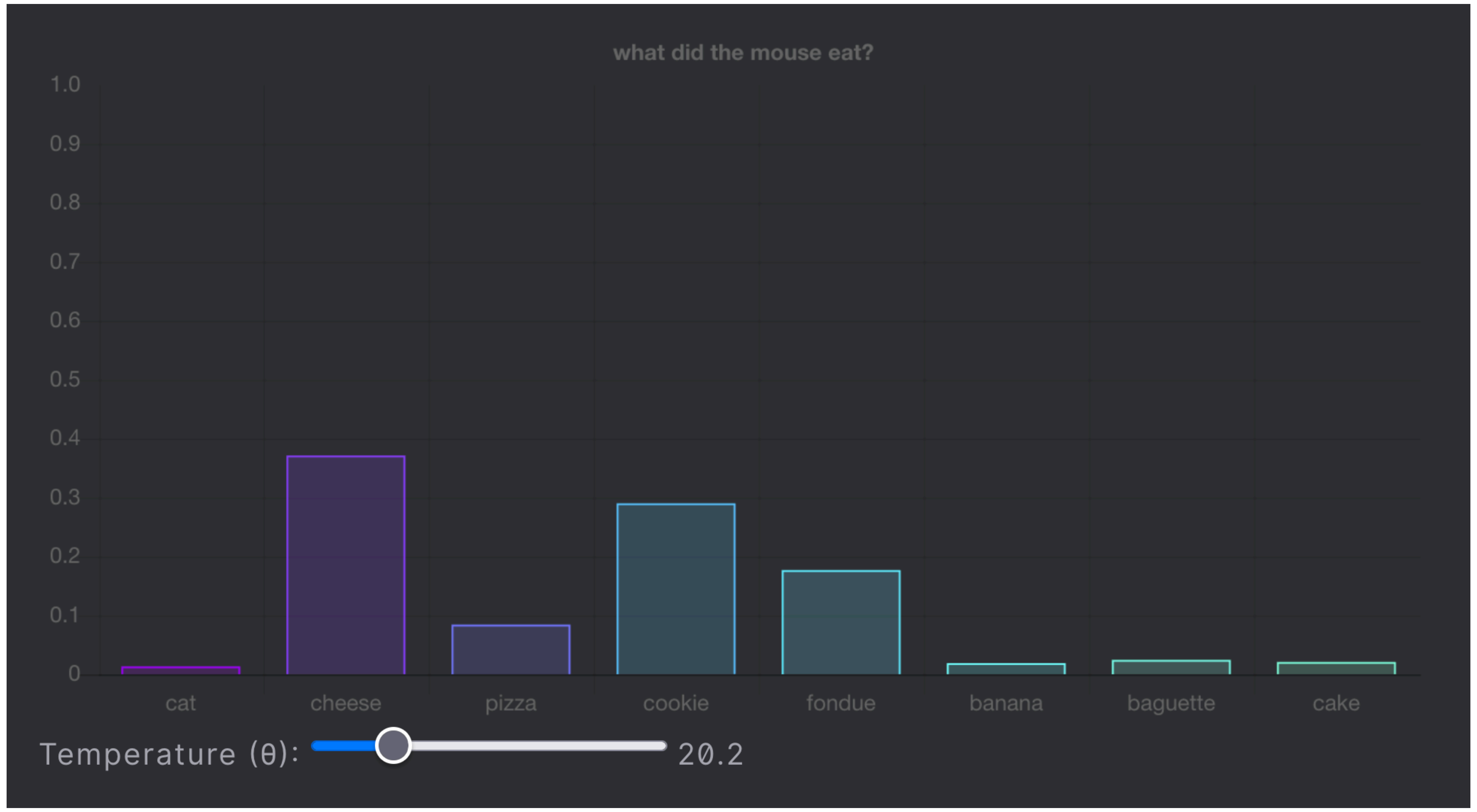
Fewer ways of 'arranging' energy amongst particles

Lower Entropy

Higher temperature
particles vibrate more

More ways of 'arranging' energy amongst particles

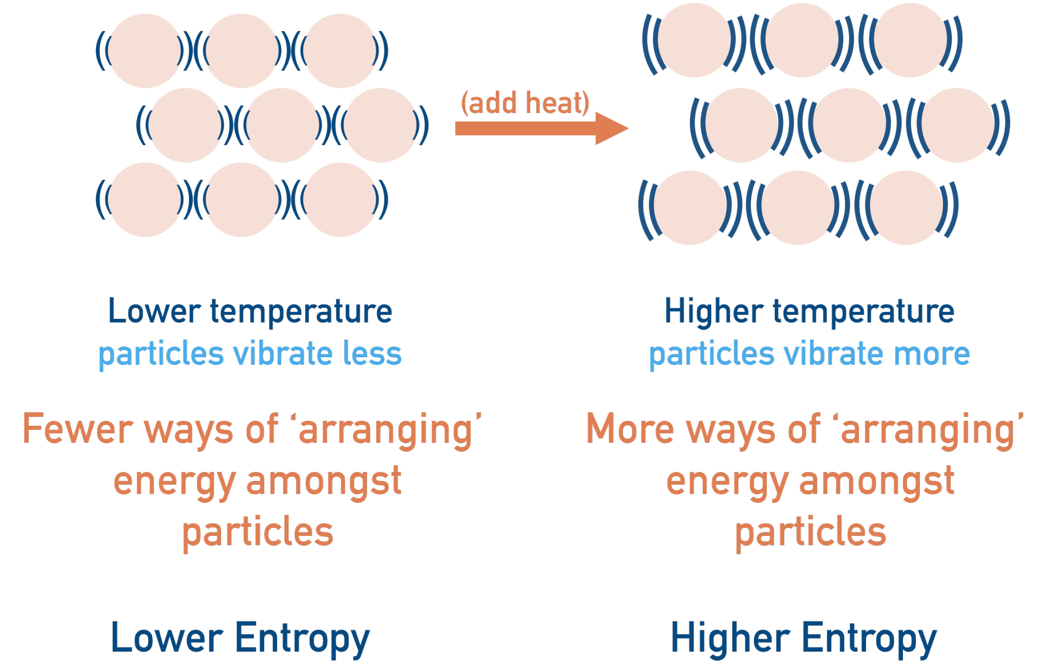
Higher Entropy



Temperature (θ): 20.2

Softmax Temperature

- Takes inspiration from temperature in **physics**
  - At low temperatures, probability **"solidifies"** around the most-probable logits



(add heat)

Lower temperature
particles vibrate less

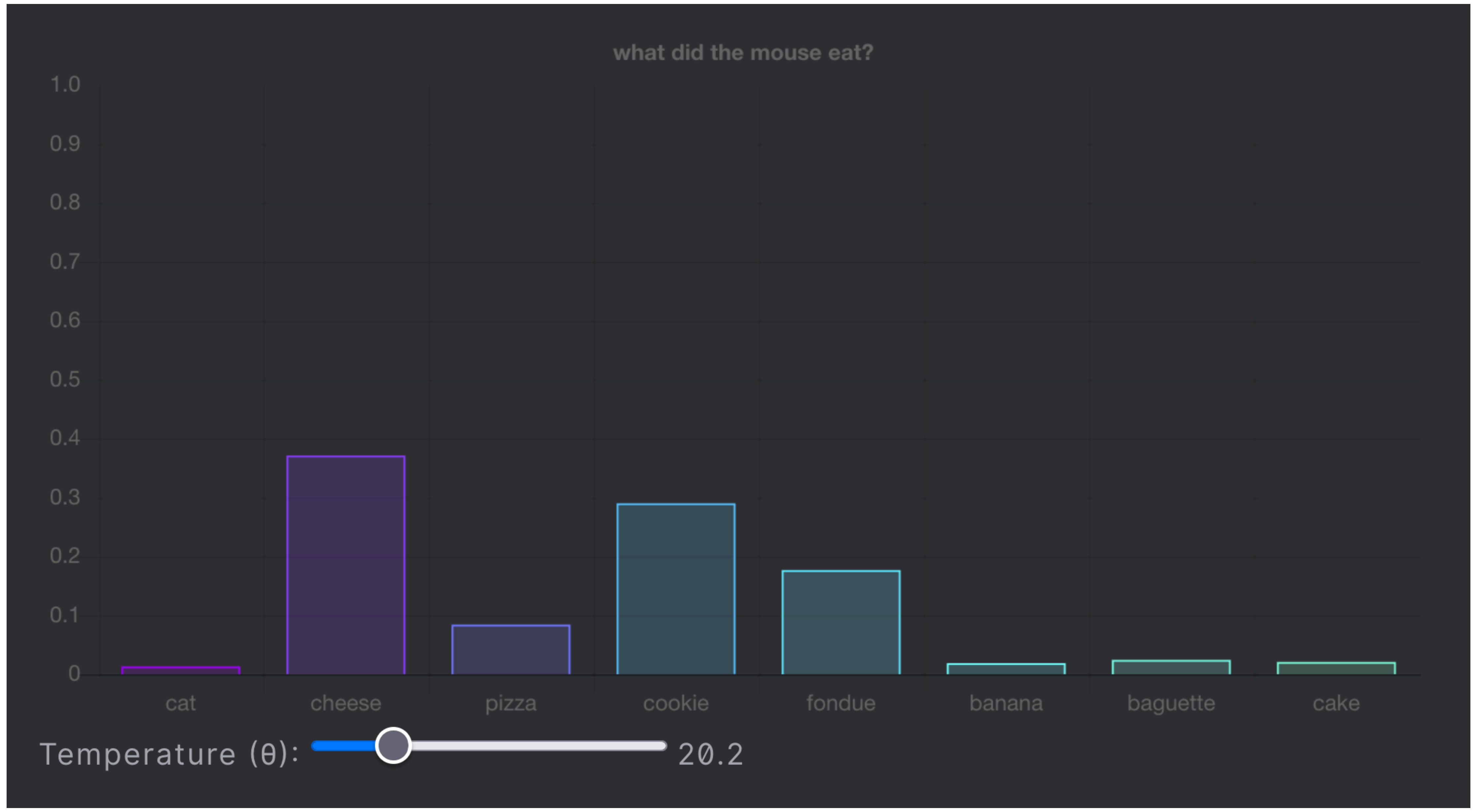
Fewer ways of 'arranging' energy amongst particles

Lower Entropy

Higher temperature
particles vibrate more

More ways of 'arranging' energy amongst particles

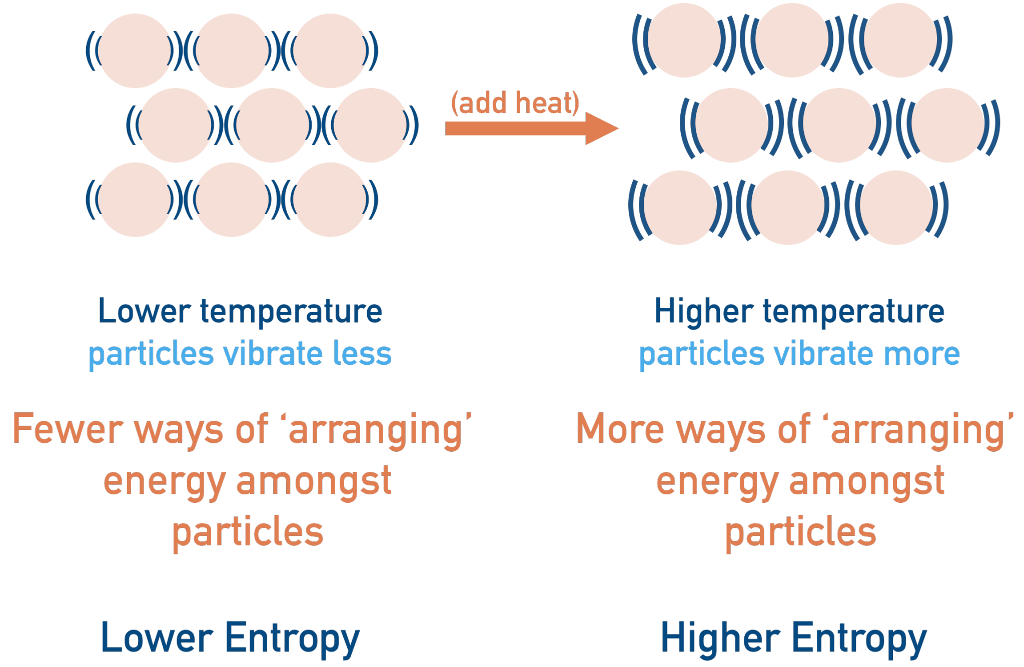
Higher Entropy

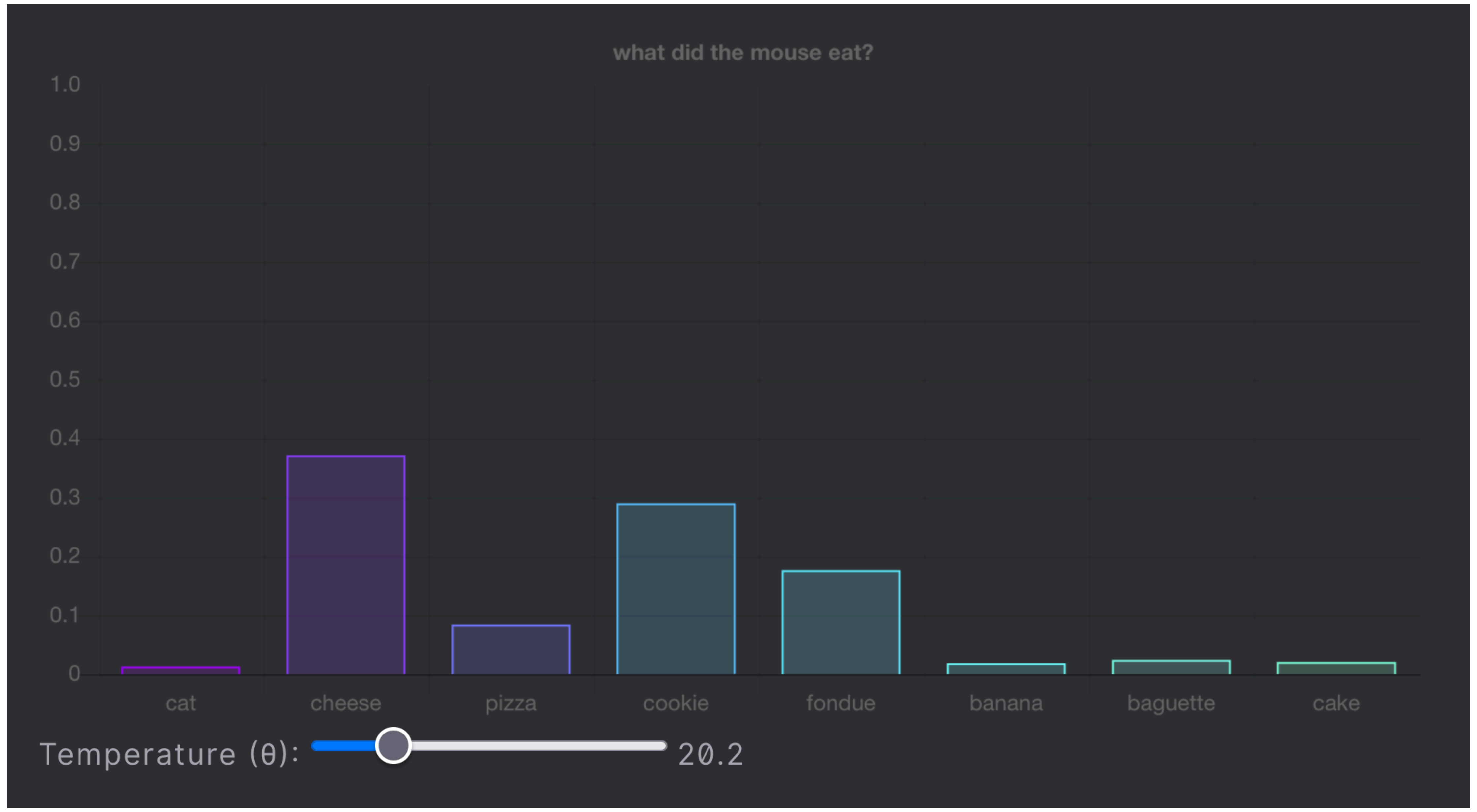


Temperature (θ): 20.2

Softmax Temperature

- Takes inspiration from temperature in **physics**
  - At low temperatures, probability **"solidifies"** around the most-probable logits
  - At high temperatures, probability acts like a "gas" and **distributes widely**

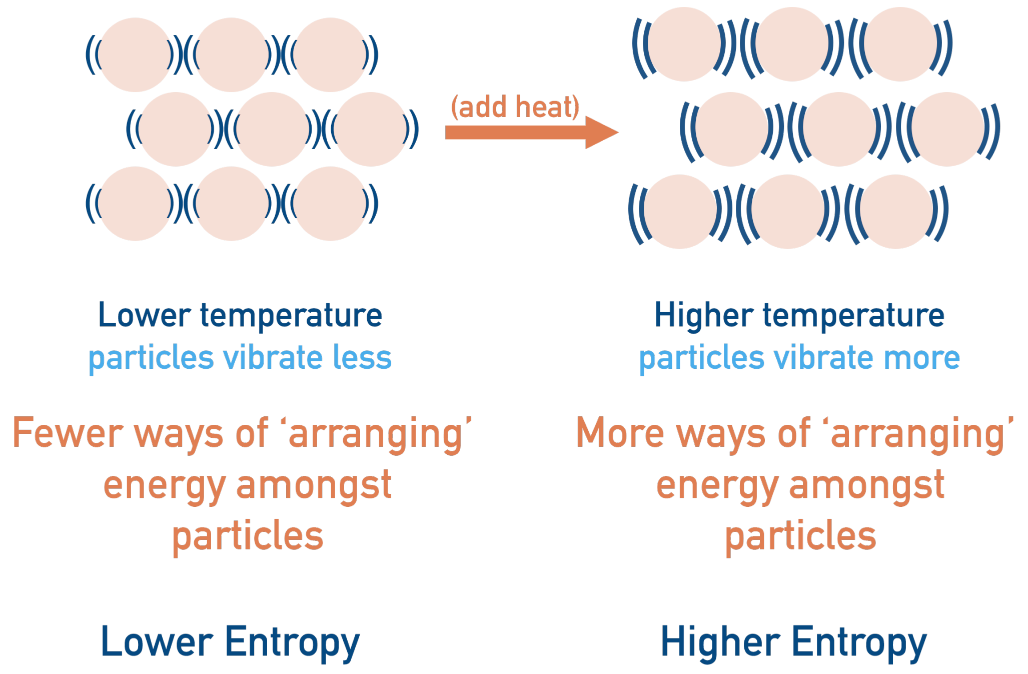




Temperature (θ): 20.2

Softmax Temperature

- Takes inspiration from temperature in **physics**
  - At low temperatures, probability **"solidifies"** around the most-probable logits
  - At high temperatures, probability acts like a "gas" and **distributes widely**
- Great visualization tool available on this blog


(add heat)

Lower temperature
particles vibrate less

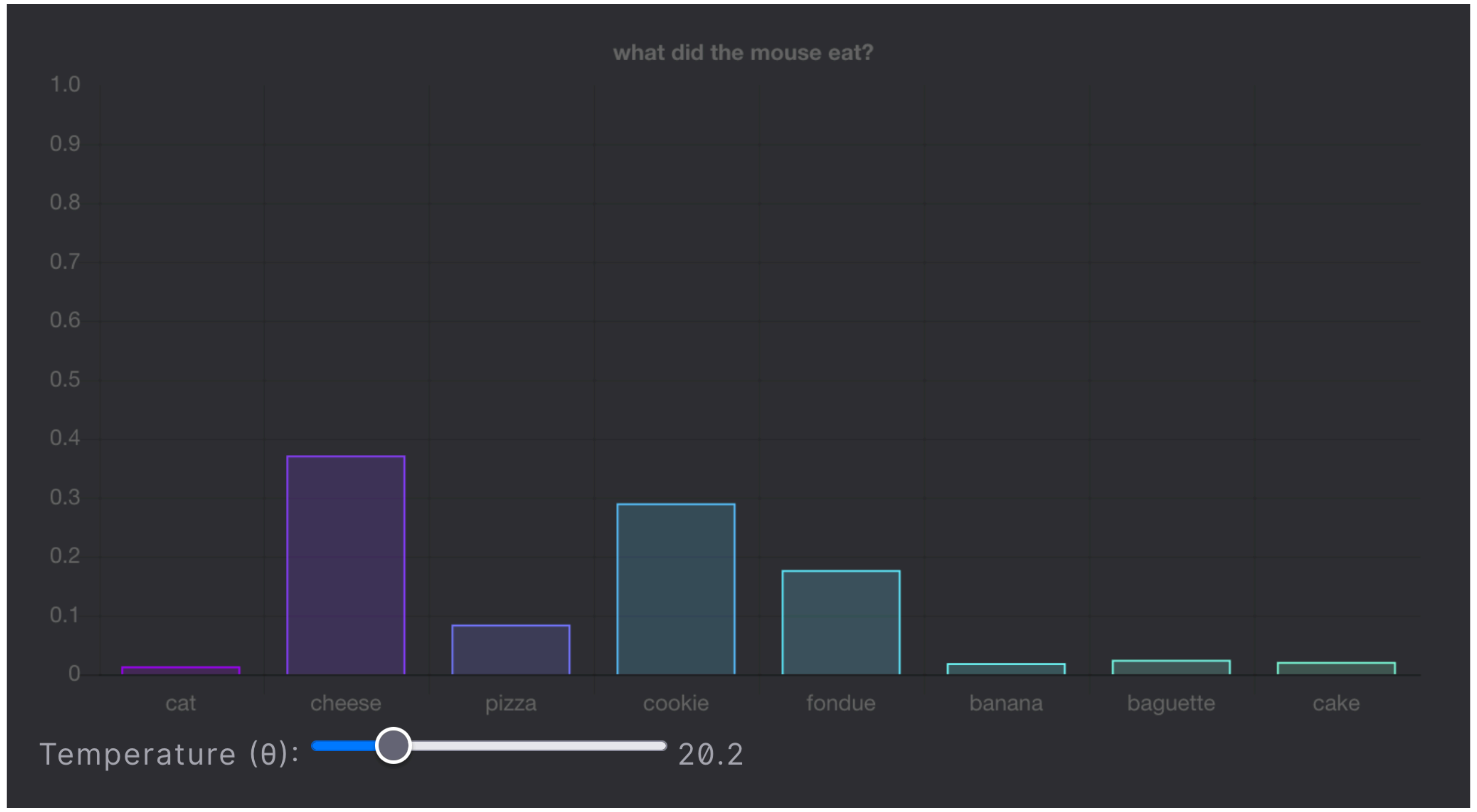
Fewer ways of 'arranging' energy amongst particles

Lower Entropy

Higher temperature
particles vibrate more

More ways of 'arranging' energy amongst particles

Higher Entropy


what did the mouse eat?

Temperature (θ): 20.2

Beam Search

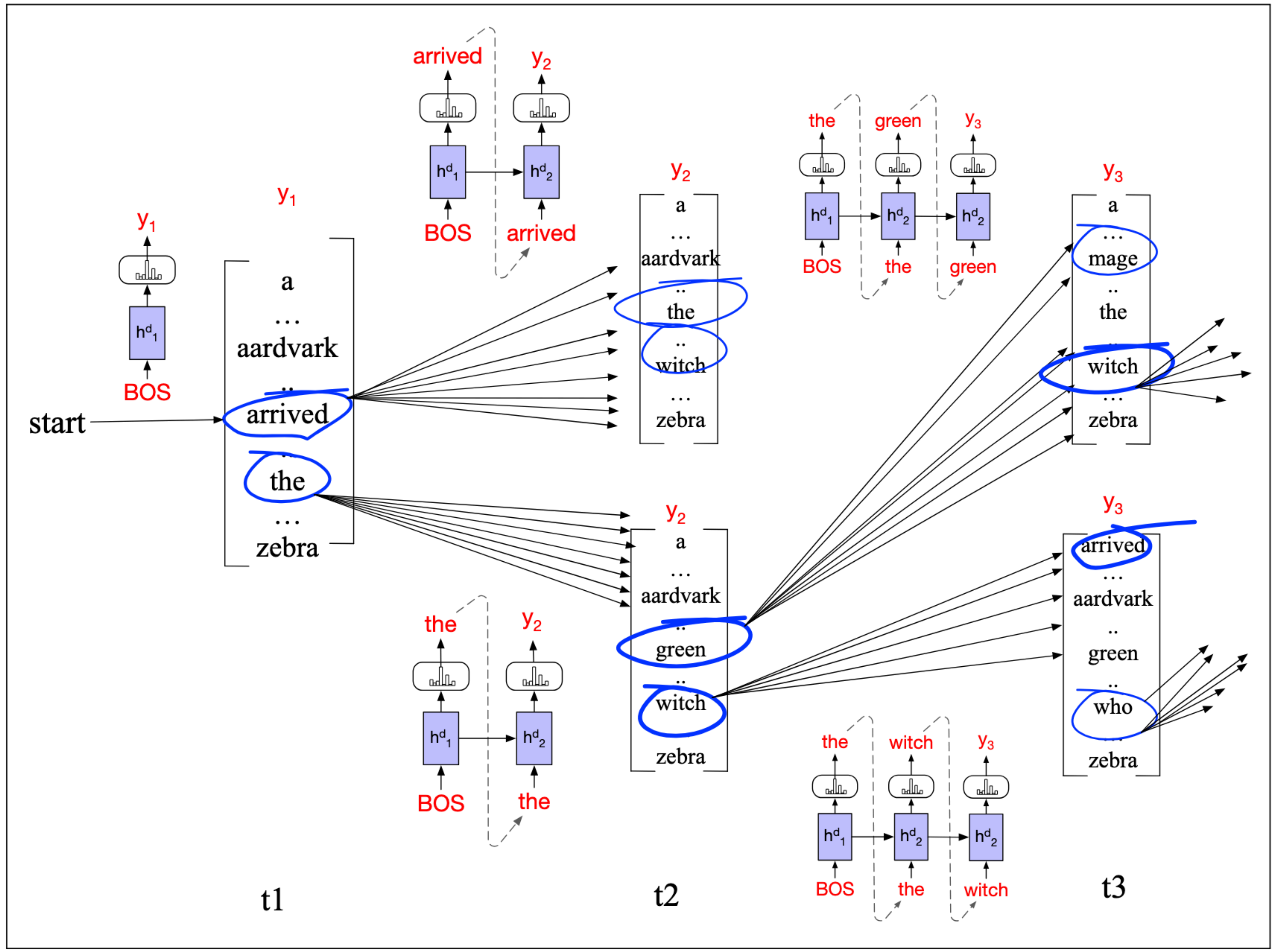


Figure 12.8 Beam search decoding with a beam width of $k = 2$. At each time step, we choose the k best hypotheses, form the V possible extensions of each, score those $k \times V$ hypotheses and choose the best $k = 2$ to continue. At time 1, the frontier has the best 2 options from the initial decoder state: *arrived* and *the*. We extend each, compute the probability of all the hypotheses so far (*arrived the*, *arrived aardvark*, *the green*, *the witch*) and again chose the best 2 (*the green* and *the witch*) to be the search frontier. The images on the arcs schematically represent the decoders that must be run at each step to score the next words (for simplicity not depicting cross-attention).

Beam Search

- Recall that greedy decoding does **not guarantee** the overall **highest-probability sequence**
    - (In fact, it probably won't be)

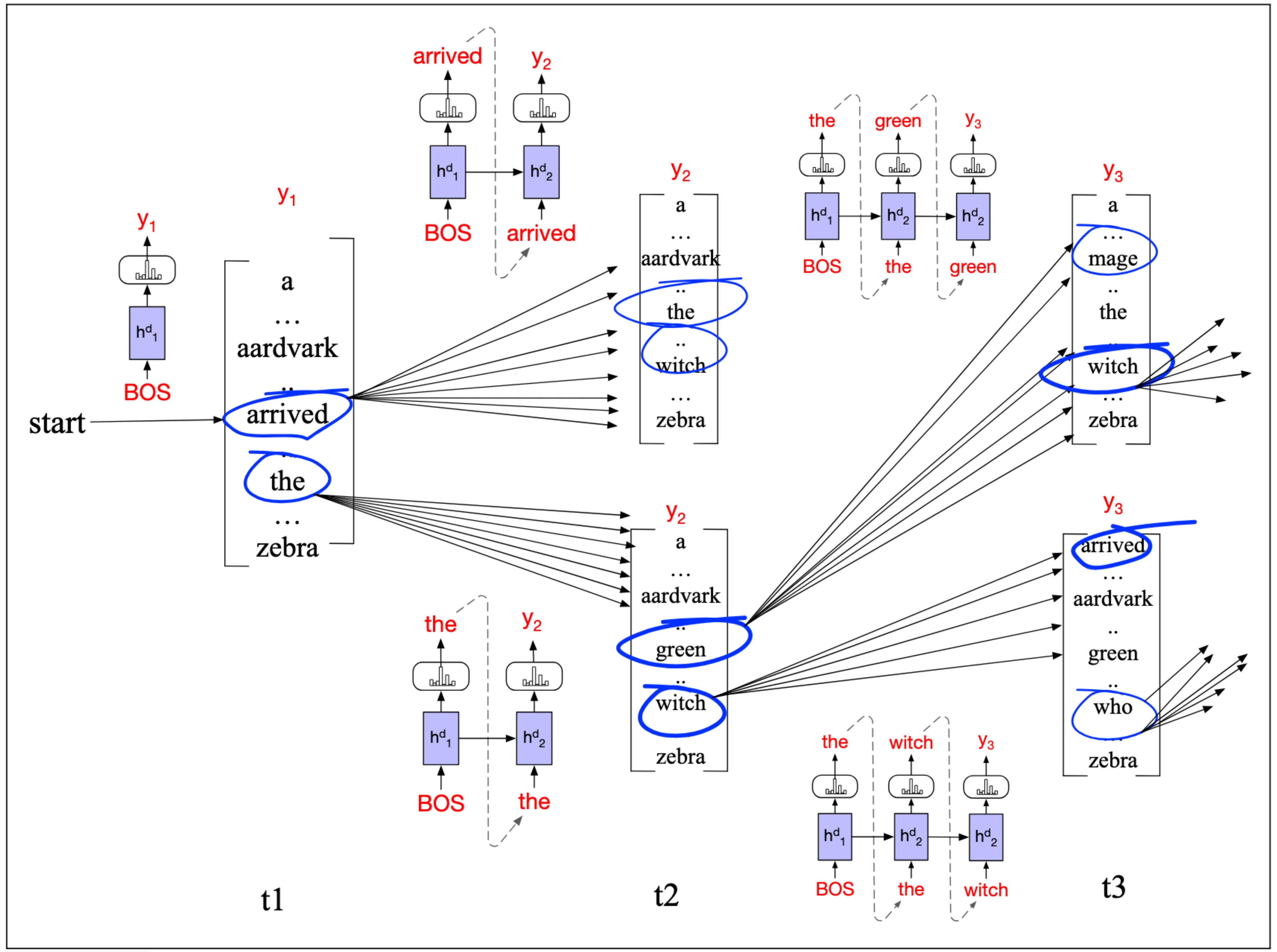


Figure 12.8 Beam search decoding with a beam width of $k = 2$. At each time step, we choose the k best hypotheses, form the V possible extensions of each, score those $k \times V$ hypotheses and choose the best $k = 2$ to continue. At time 1, the frontier has the best 2 options from the initial decoder state: *arrived* and *the*. We extend each, compute the probability of all the hypotheses so far (*arrived the*, *arrived aardvark*, *the green*, *the witch*) and again chose the best 2 (*the green* and *the witch*) to be the search frontier. The images on the arcs schematically represent the decoders that must be run at each step to score the next words (for simplicity not depicting cross-attention).

Beam Search

- Recall that greedy decoding does **not guarantee** the overall **highest-probability sequence**
  - (In fact, it probably won't be)
- The space of **all possible sequences** is massive! $|V|^N$
  - We need a better way to **search** for the optimum sequence

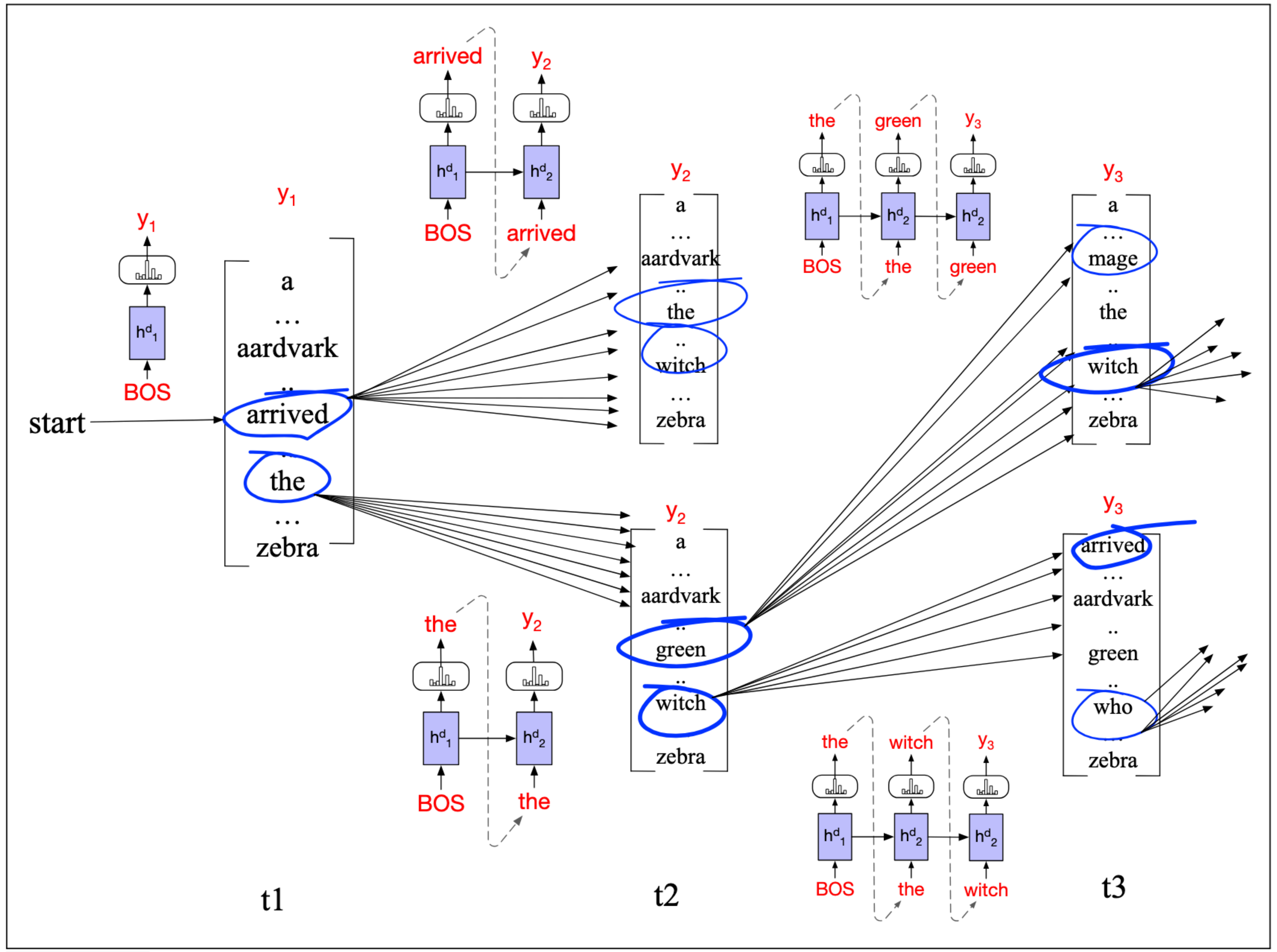


Figure 12.8 Beam search decoding with a beam width of $k = 2$. At each time step, we choose the k best hypotheses, form the V possible extensions of each, score those $k \times V$ hypotheses and choose the best $k = 2$ to continue. At time 1, the frontier has the best 2 options from the initial decoder state: *arrived* and *the*. We extend each, compute the probability of all the hypotheses so far (*arrived the*, *arrived aardvark*, *the green*, *the witch*) and again chose the best 2 (*the green* and *the witch*) to be the search frontier. The images on the arcs schematically represent the decoders that must be run at each step to score the next words (for simplicity not depicting cross-attention).

Beam Search

- Recall that greedy decoding does **not guarantee** the overall **highest-probability sequence**
  - (In fact, it probably won't be)
- The space of **all possible sequences** is massive! $|V|^N$
  - We need a better way to **search** for the optimum sequence
- **Beam Search**: at each step, choose the **top-k most-probable** continuations
  - Always keep the **k most-probable paths** in contention, and **prune** others
  - These paths often called "beams"

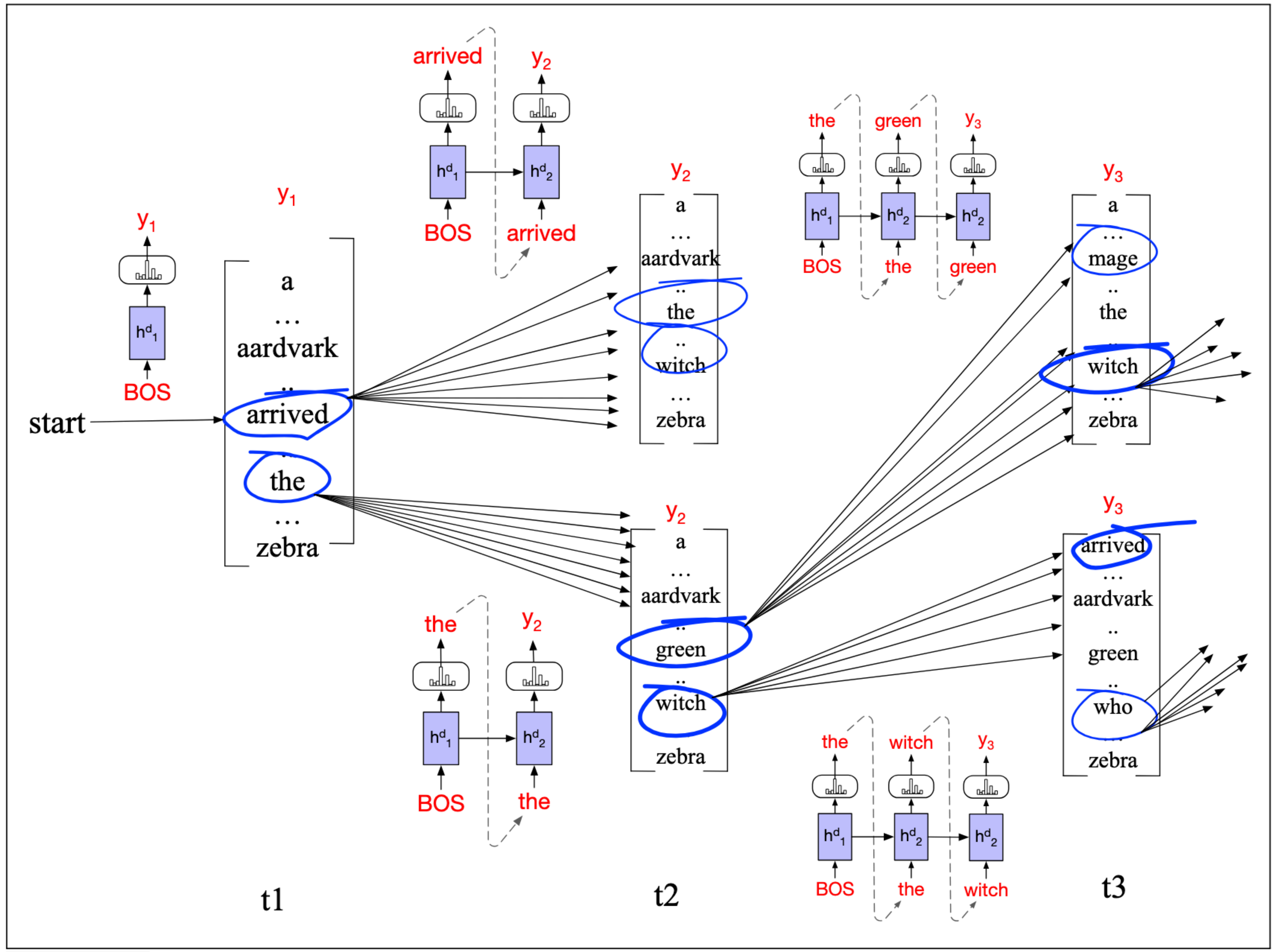


Figure 12.8 Beam search decoding with a beam width of $k = 2$. At each time step, we choose the k best hypotheses, form the V possible extensions of each, score those $k \times V$ hypotheses and choose the best $k = 2$ to continue. At time 1, the frontier has the best 2 options from the initial decoder state: *arrived* and *the*. We extend each, compute the probability of all the hypotheses so far (*arrived the*, *arrived aardvark*, *the green*, *the witch*) and again chose the best 2 (*the green* and *the witch*) to be the search frontier. The images on the arcs schematically represent the decoders that must be run at each step to score the next words (for simplicity not depicting cross-attention).

Beam Search

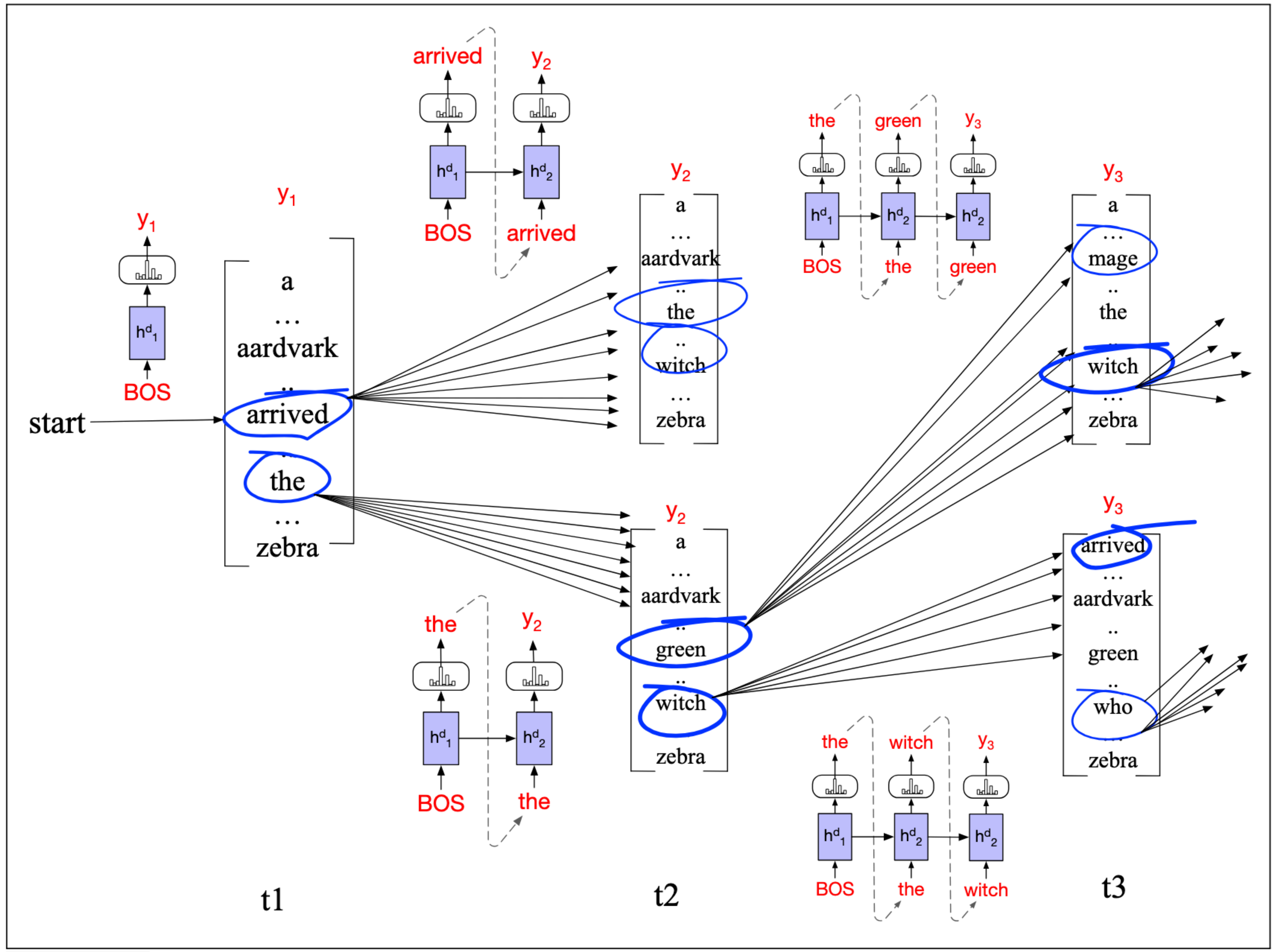


Figure 12.8 Beam search decoding with a beam width of $k = 2$. At each time step, we choose the k best hypotheses, form the V possible extensions of each, score those $k \times V$ hypotheses and choose the best $k = 2$ to continue. At time 1, the frontier has the best 2 options from the initial decoder state: *arrived* and *the*. We extend each, compute the probability of all the hypotheses so far (*arrived the*, *arrived aardvark*, *the green*, *the witch*) and again chose the best 2 (*the green* and *the witch*) to be the search frontier. The images on the arcs schematically represent the decoders that must be run at each step to score the next words (for simplicity not depicting cross-attention).

Beam Search

- In practice, Beam Search is only used for **particular NLP applications**

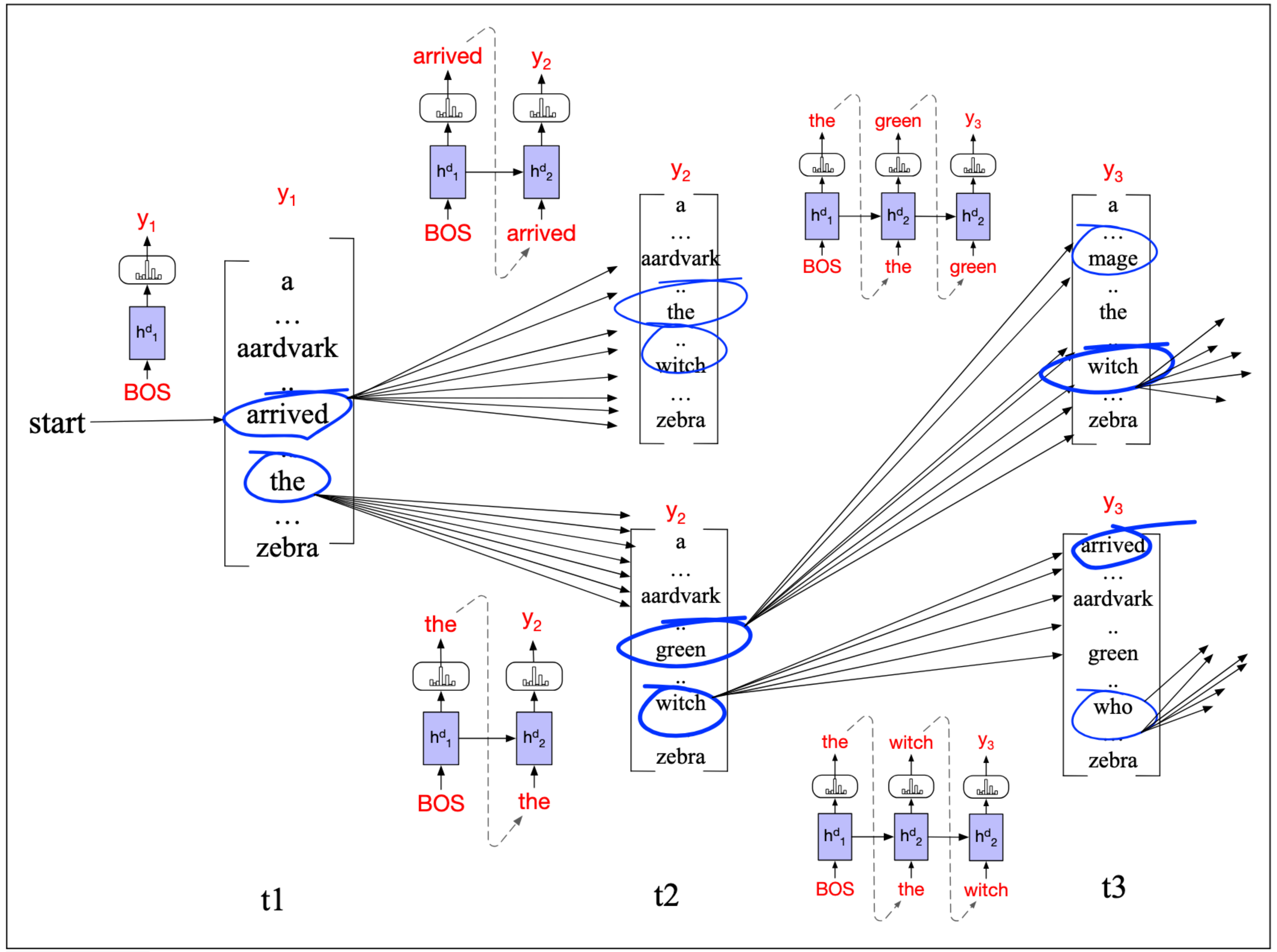


Figure 12.8 Beam search decoding with a beam width of $k = 2$. At each time step, we choose the k best hypotheses, form the V possible extensions of each, score those $k \times V$ hypotheses and choose the best $k = 2$ to continue. At time 1, the frontier has the best 2 options from the initial decoder state: *arrived* and *the*. We extend each, compute the probability of all the hypotheses so far (*arrived the*, *arrived aardvark*, *the green*, *the witch*) and again chose the best 2 (*the green* and *the witch*) to be the search frontier. The images on the arcs schematically represent the decoders that must be run at each step to score the next words (for simplicity not depicting cross-attention).

Beam Search

- In practice, Beam Search is only used for **particular NLP applications**
  - Recall that we **might not want** the most probable sequence (often boring/ memorized)

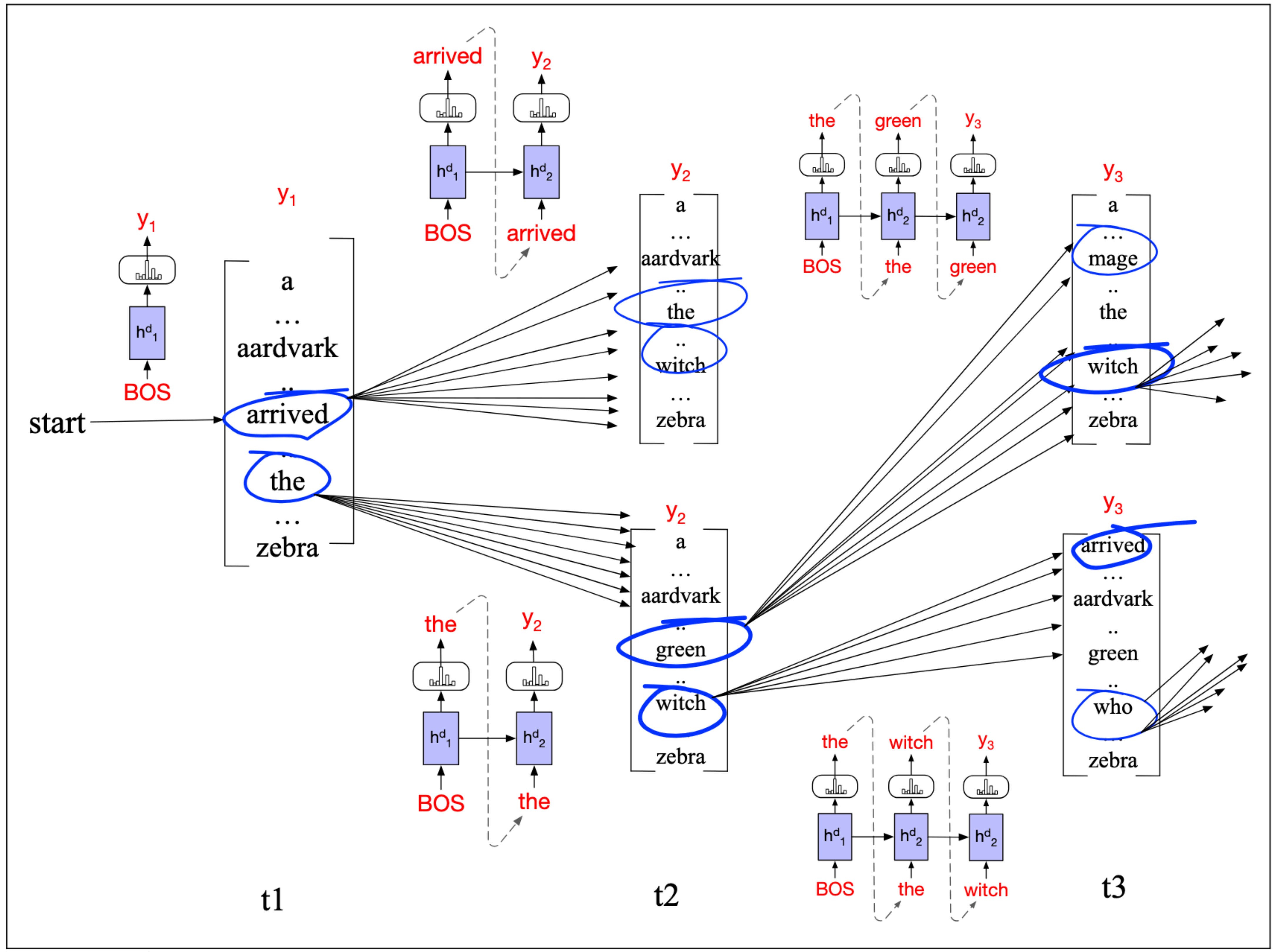


Figure 12.8 Beam search decoding with a beam width of $k = 2$. At each time step, we choose the k best hypotheses, form the V possible extensions of each, score those $k \times V$ hypotheses and choose the best $k = 2$ to continue. At time 1, the frontier has the best 2 options from the initial decoder state: *arrived* and *the*. We extend each, compute the probability of all the hypotheses so far (*arrived the*, *arrived aardvark*, *the green*, *the witch*) and again chose the best 2 (*the green* and *the witch*) to be the search frontier. The images on the arcs schematically represent the decoders that must be run at each step to score the next words (for simplicity not depicting cross-attention).

Beam Search

- In practice, Beam Search is only used for **particular NLP applications**
  - Recall that we **might not want** the most probable sequence (often boring/ memorized)
  - Beam Search is used where there's emphasis on a **correct answer** (e.g. Machine Translation)

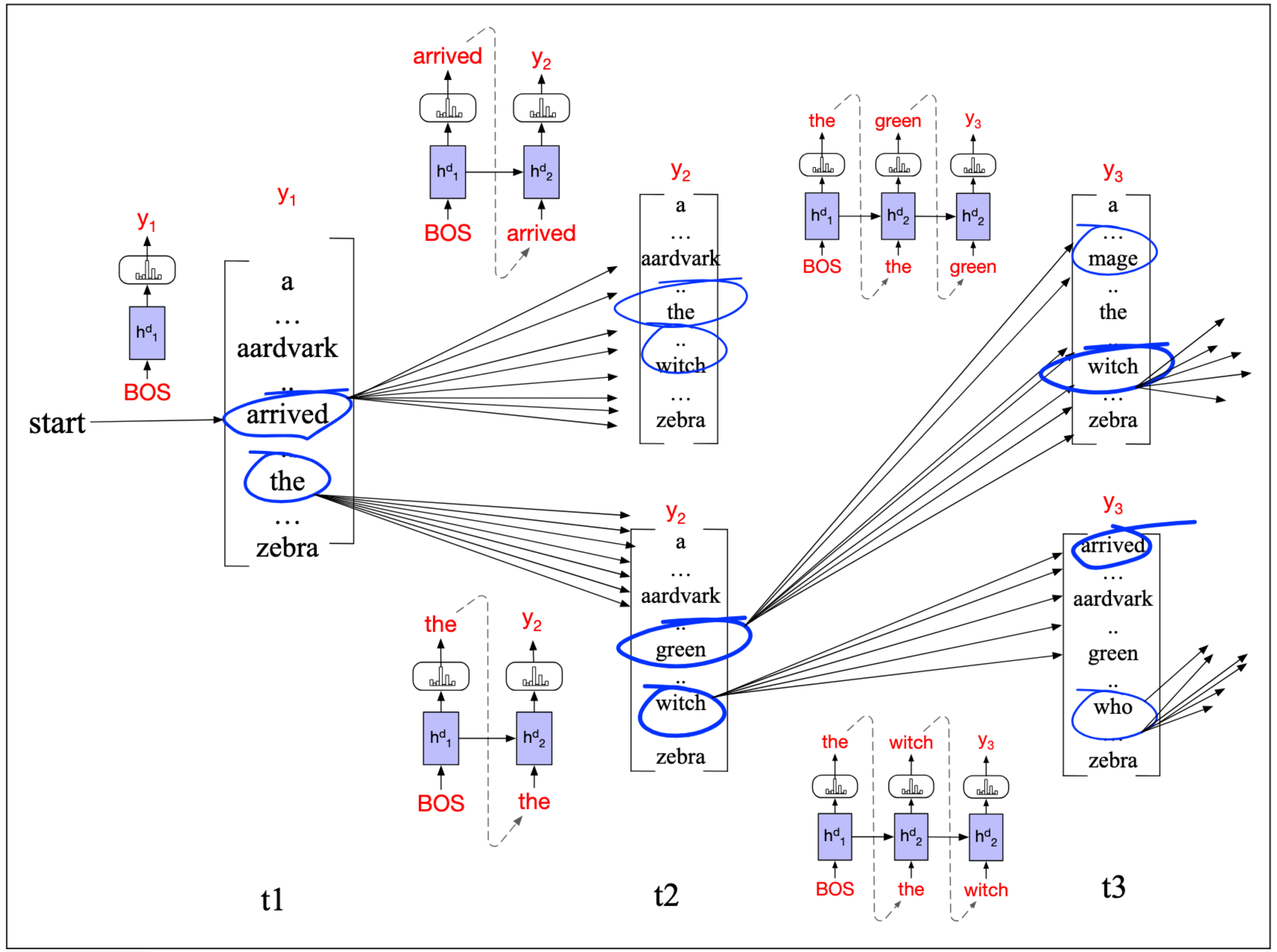


Figure 12.8 Beam search decoding with a beam width of $k = 2$. At each time step, we choose the k best hypotheses, form the V possible extensions of each, score those $k \times V$ hypotheses and choose the best $k = 2$ to continue. At time 1, the frontier has the best 2 options from the initial decoder state: *arrived* and *the*. We extend each, compute the probability of all the hypotheses so far (*arrived the*, *arrived aardvark*, *the green*, *the witch*) and again chose the best 2 (*the green* and *the witch*) to be the search frontier. The images on the arcs schematically represent the decoders that must be run at each step to score the next words (for simplicity not depicting cross-attention).

Beam Search

- In practice, Beam Search is only used for **particular NLP applications**
  - Recall that we **might not want** the most probable sequence (often boring/ memorized)
  - Beam Search is used where there's emphasis on a **correct answer** (e.g. Machine Translation)
- Fairly **computationally expensive** compared to other decoding methods

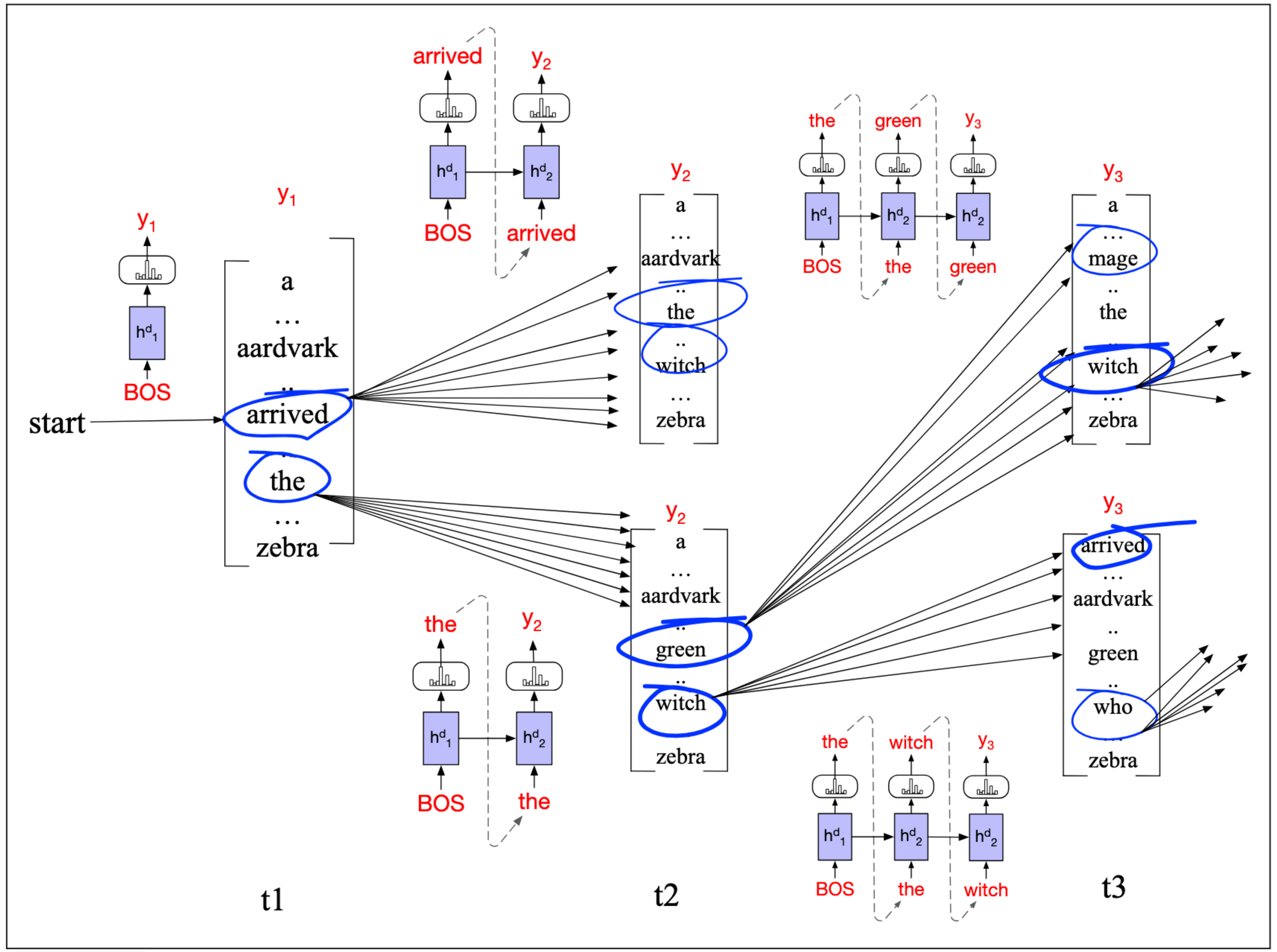
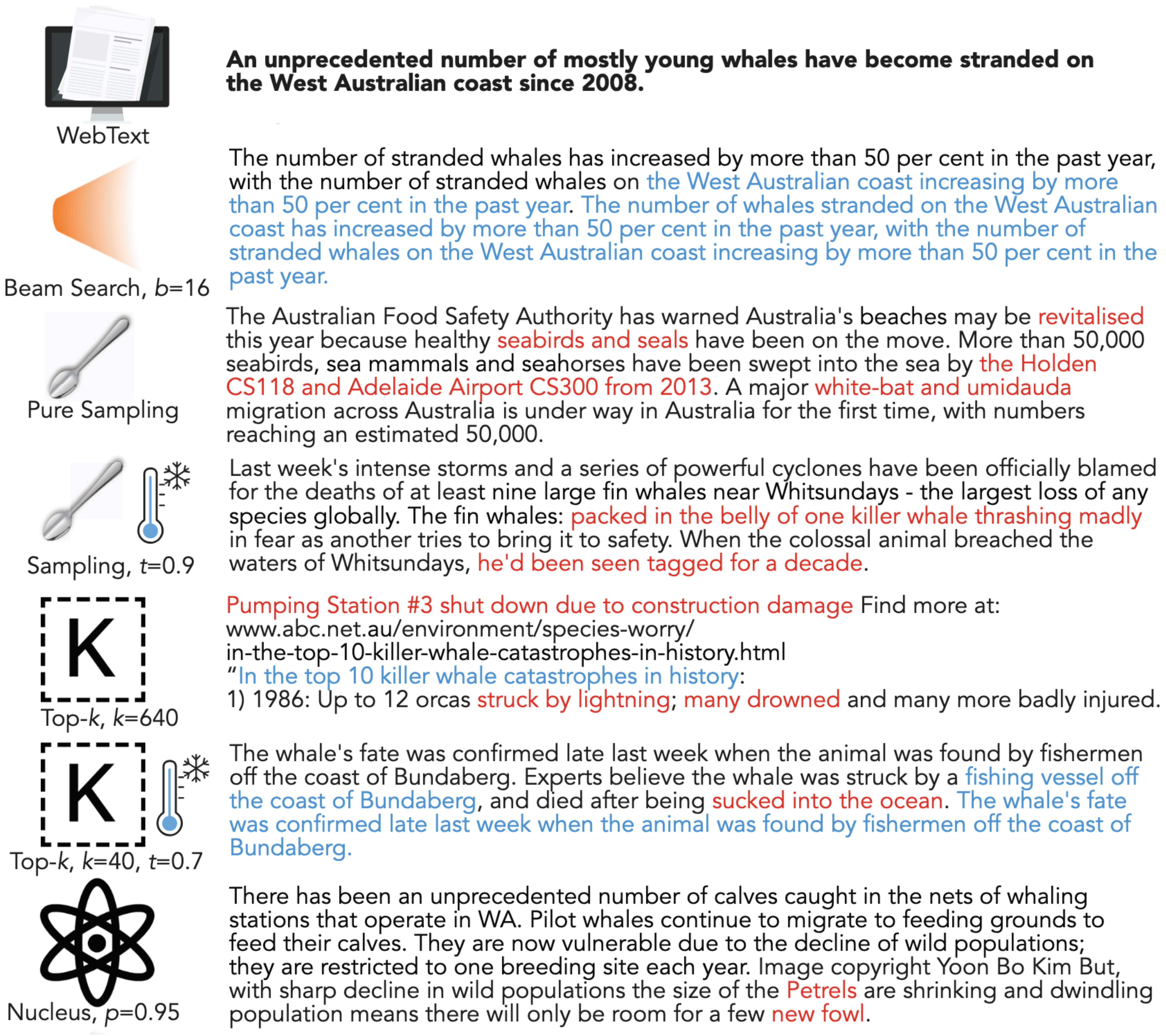


Figure 12.8 Beam search decoding with a beam width of $k = 2$. At each time step, we choose the k best hypotheses, form the V possible extensions of each, score those $k \times V$ hypotheses and choose the best $k = 2$ to continue. At time 1, the frontier has the best 2 options from the initial decoder state: *arrived* and *the*. We extend each, compute the probability of all the hypotheses so far (*arrived the*, *arrived aardvark*, *the green*, *the witch*) and again chose the best 2 (*the green* and *the witch*) to be the search frontier. The images on the arcs schematically represent the decoders that must be run at each step to score the next words (for simplicity not depicting cross-attention).

Generation Big Picture



WebText

An unprecedented number of mostly young whales have become stranded on the West Australian coast since 2008.

Beam Search, *b*=16

The number of stranded whales has increased by more than 50 per cent in the past year, with the number of stranded whales on the West Australian coast increasing by more than 50 per cent in the past year. The number of whales stranded on the West Australian coast has increased by more than 50 per cent in the past year, with the number of stranded whales on the West Australian coast increasing by more than 50 per cent in the past year.

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.

Top-*k*, *k*=640

Pumping Station #3 shut down due to construction damage Find more at: www.abc.net.au/environment/species-worry/in-the-top-10-killer-whale-catastrophes-in-history.html
“In the top 10 killer whale catastrophes in history:
1) 1986: Up to 12 orcas struck by lightning; many drowned and many more badly injured.

Top-*k*, *k*=40, *t*=0.7

The whale's fate was confirmed late last week when the animal was found by fishermen off the coast of Bundaberg. Experts believe the whale was struck by a fishing vessel off the coast of Bundaberg, and died after being sucked into the ocean. The whale's fate was confirmed late last week when the animal was found by fishermen off the coast of Bundaberg.

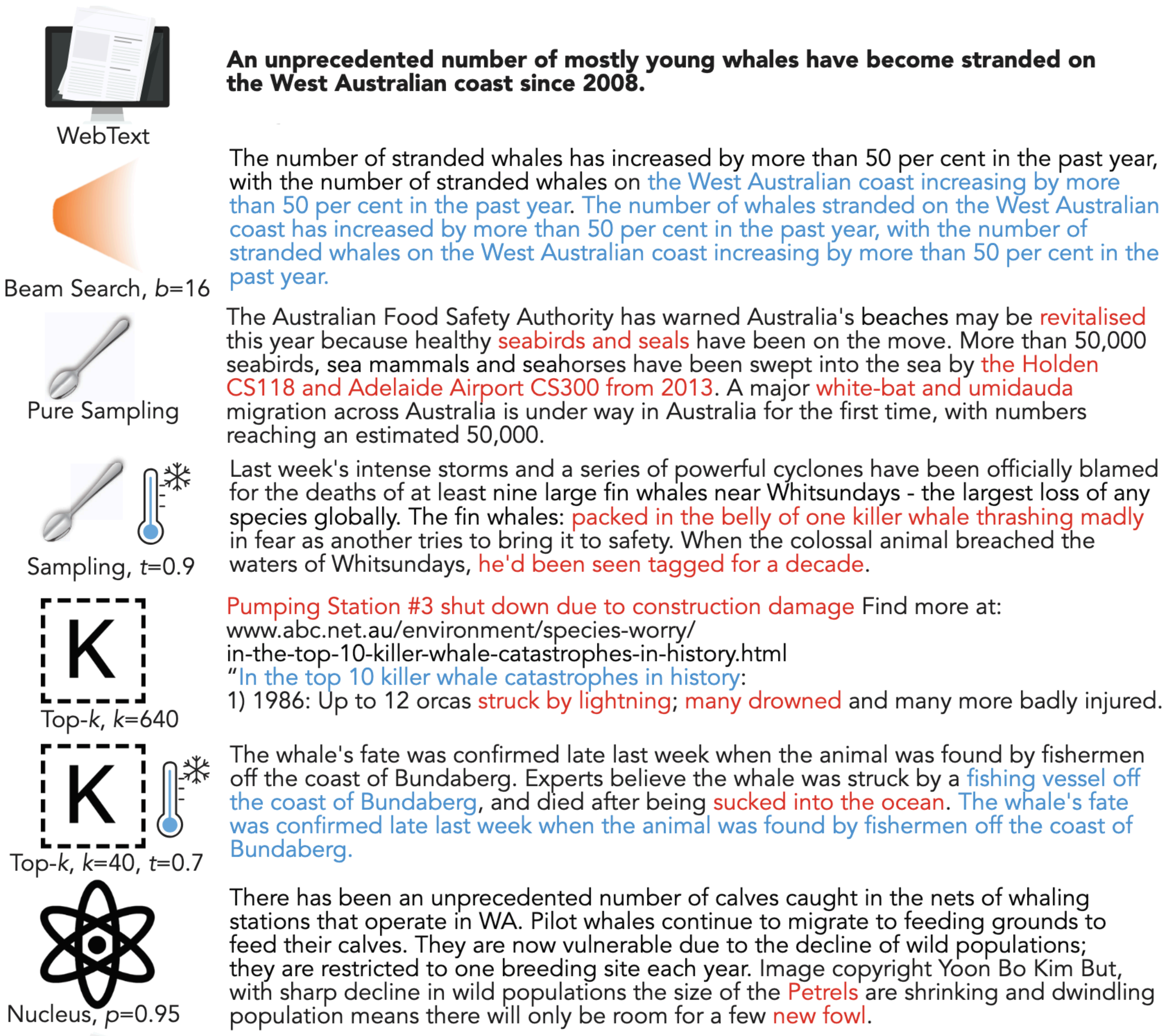
Nucleus, *p*=0.95

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.

Holtzman et al (2020)

Generation Big Picture

- In practice, we often **combine generation techniques**



WebText

An unprecedented number of mostly young whales have become stranded on the West Australian coast since 2008.

Beam Search, *b*=16

The number of stranded whales has increased by more than 50 per cent in the past year, with the number of stranded whales on the West Australian coast increasing by more than 50 per cent in the past year. The number of whales stranded on the West Australian coast has increased by more than 50 per cent in the past year, with the number of stranded whales on the West Australian coast increasing by more than 50 per cent in the past year.

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.

Top-*k*, *k*=640

Pumping Station #3 shut down due to construction damage Find more at: www.abc.net.au/environment/species-worry/in-the-top-10-killer-whale-catastrophes-in-history.html
"In the top 10 killer whale catastrophes in history:
1) 1986: Up to 12 orcas struck by lightning; many drowned and many more badly injured.

Top-*k*, *k*=40, *t*=0.7

The whale's fate was confirmed late last week when the animal was found by fishermen off the coast of Bundaberg. Experts believe the whale was struck by a fishing vessel off the coast of Bundaberg, and died after being sucked into the ocean. The whale's fate was confirmed late last week when the animal was found by fishermen off the coast of Bundaberg.

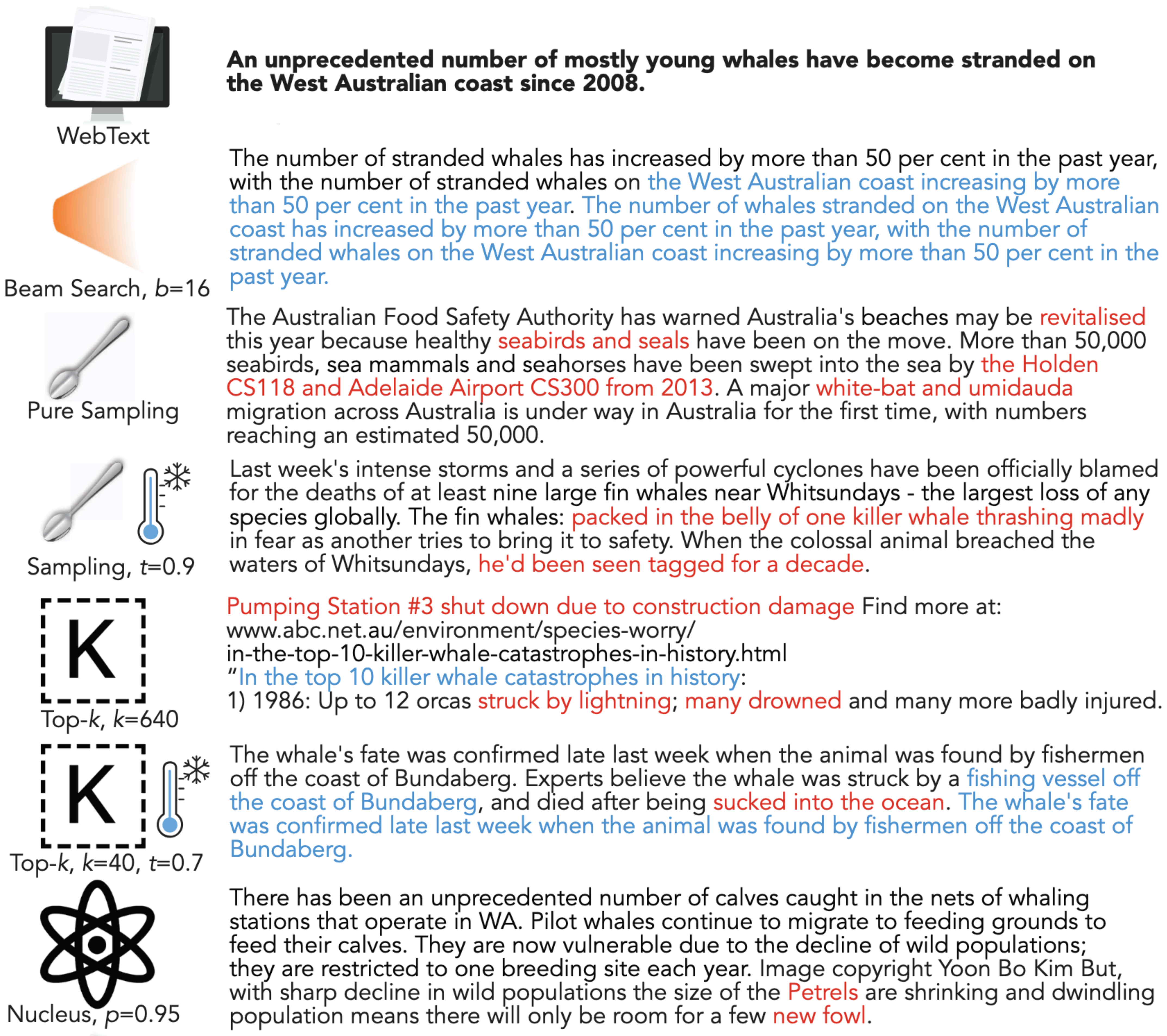
Nucleus, *p*=0.95

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.

Holtzman et al (2020)

Generation Big Picture

- In practice, we often **combine generation techniques**
  - E.g. top-k and temperature



WebText

An unprecedented number of mostly young whales have become stranded on the West Australian coast since 2008.

Beam Search, *b*=16

The number of stranded whales has increased by more than 50 per cent in the past year, with the number of stranded whales on the West Australian coast increasing by more than 50 per cent in the past year. The number of whales stranded on the West Australian coast has increased by more than 50 per cent in the past year, with the number of stranded whales on the West Australian coast increasing by more than 50 per cent in the past year.

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.

Top-*k*, *k*=640

Pumping Station #3 shut down due to construction damage Find more at: www.abc.net.au/environment/species-worry/in-the-top-10-killer-whale-catastrophes-in-history.html
“In the top 10 killer whale catastrophes in history:
1) 1986: Up to 12 orcas struck by lightning; many drowned and many more badly injured.

Top-*k*, *k*=40, *t*=0.7

The whale's fate was confirmed late last week when the animal was found by fishermen off the coast of Bundaberg. Experts believe the whale was struck by a fishing vessel off the coast of Bundaberg, and died after being sucked into the ocean. The whale's fate was confirmed late last week when the animal was found by fishermen off the coast of Bundaberg.

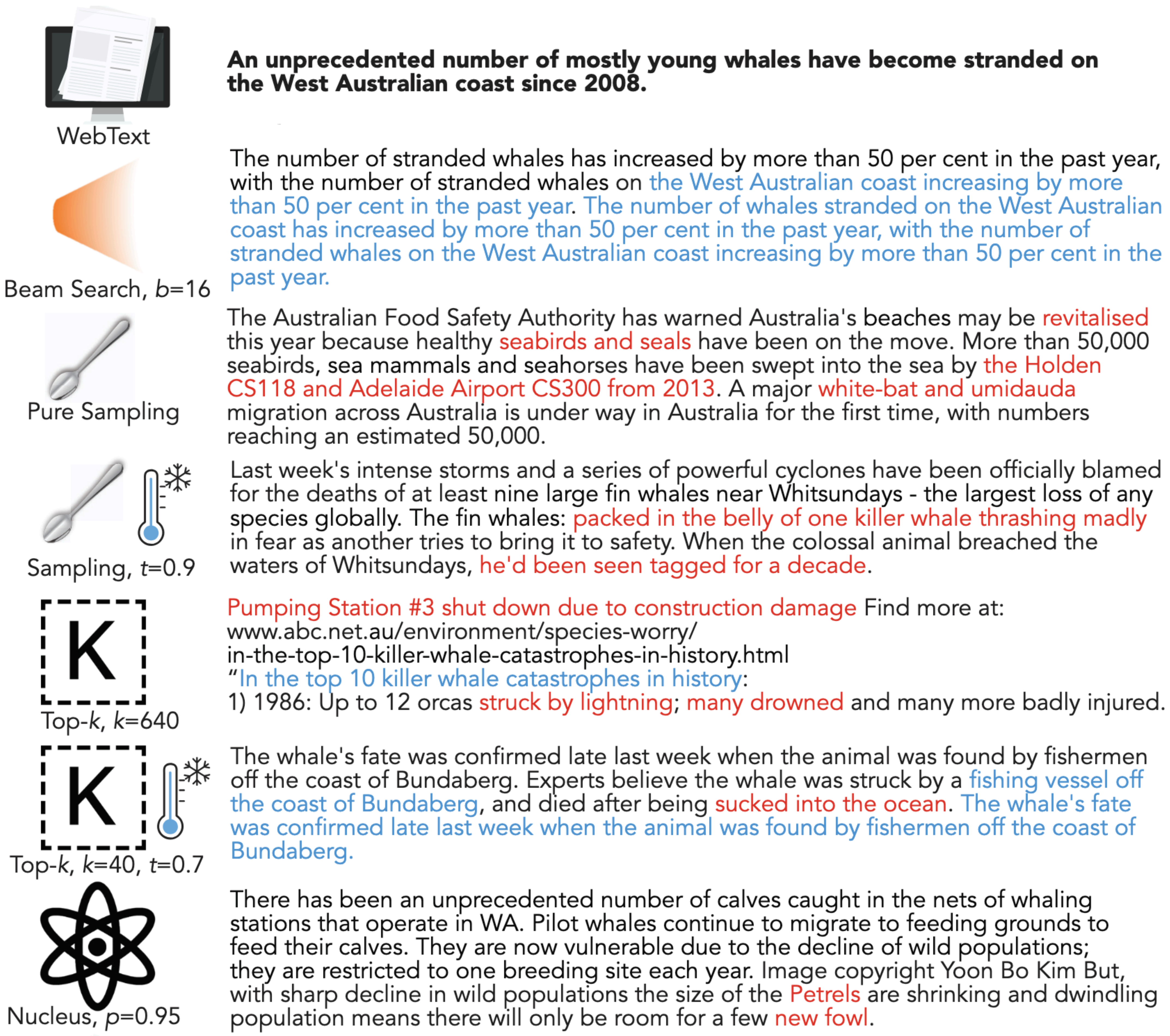
Nucleus, *p*=0.95

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.

Holtzman et al (2020)

Generation Big Picture

- In practice, we often **combine generation techniques**
  - E.g. top-k and temperature
- They have **differing drawbacks**


WebText
Beam Search, *b*=16
Pure Sampling
Sampling, *t*=0.9
K
Top-*k*, *k*=640
K
Top-*k*, *k*=40, *t*=0.7
Nucleus, *p*=0.95

An unprecedented number of mostly young whales have become stranded on the West Australian coast since 2008.

The number of stranded whales has increased by more than 50 per cent in the past year, with the number of stranded whales on the West Australian coast increasing by more than 50 per cent in the past year. The number of whales stranded on the West Australian coast has increased by more than 50 per cent in the past year, with the number of stranded whales on the West Australian coast increasing by more than 50 per cent in the past year.

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.

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.

Pumping Station #3 shut down due to construction damage Find more at: www.abc.net.au/environment/species-worry/ in-the-top-10-killer-whale-catastrophes-in-history.html
"In the top 10 killer whale catastrophes in history:
1) 1986: Up to 12 orcas struck by lightning; many drowned and many more badly injured.

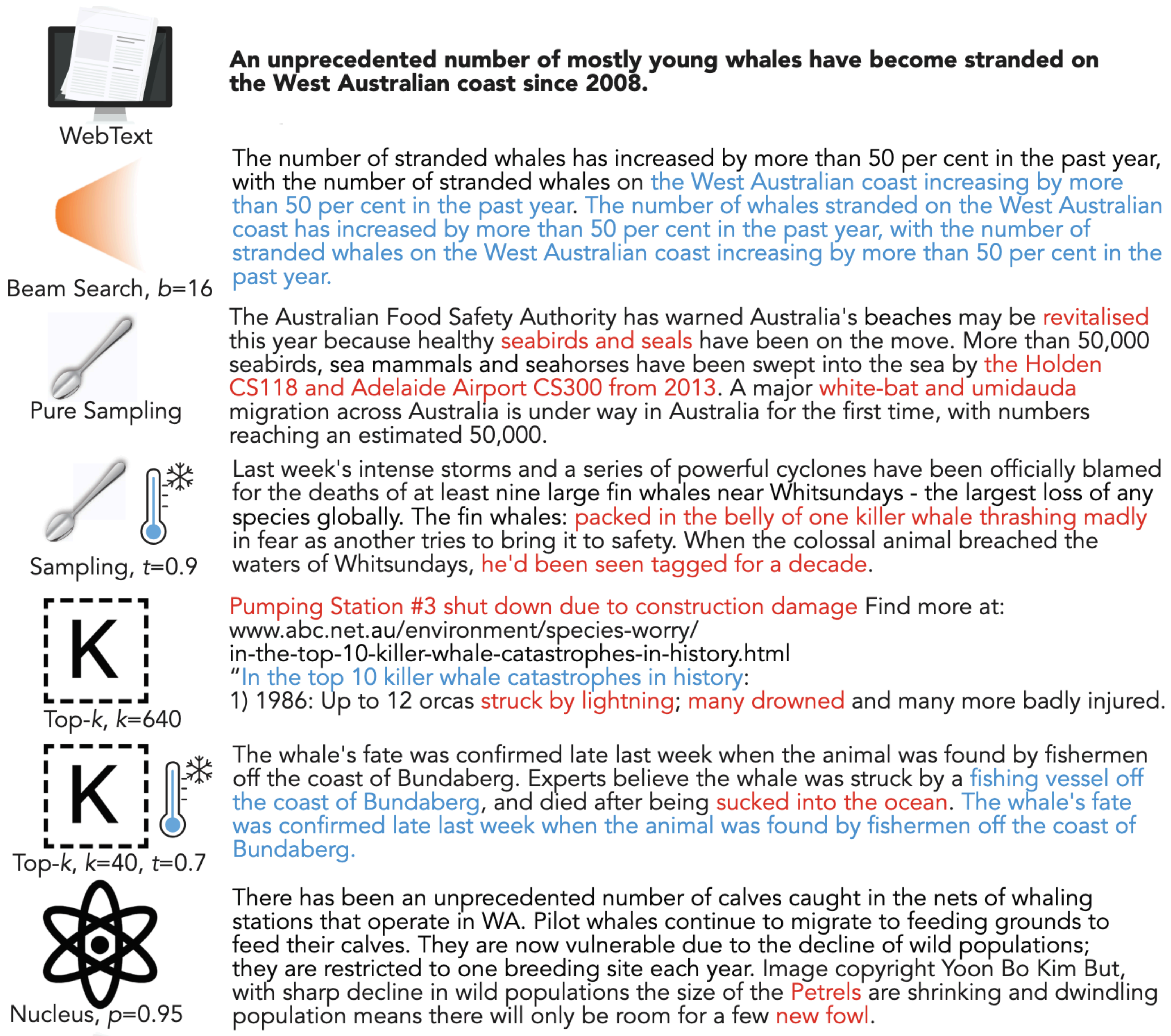
The whale's fate was confirmed late last week when the animal was found by fishermen off the coast of Bundaberg. Experts believe the whale was struck by a fishing vessel off the coast of Bundaberg, and died after being sucked into the ocean. The whale's fate was confirmed late last week when the animal was found by fishermen off the coast of Bundaberg.

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.

Holtzman et al (2020)

Generation Big Picture

- In practice, we often **combine generation techniques**
  - E.g. top-k and temperature
- They have **differing drawbacks**
  - Greedy, Top-k, and Beam Search often lead to **repetitive** generation (blue)



WebText

An unprecedented number of mostly young whales have become stranded on the West Australian coast since 2008.

Beam Search, *b*=16

The number of stranded whales has increased by more than 50 per cent in the past year, with the number of stranded whales on the West Australian coast increasing by more than 50 per cent in the past year. The number of whales stranded on the West Australian coast has increased by more than 50 per cent in the past year, with the number of stranded whales on the West Australian coast increasing by more than 50 per cent in the past year.

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.

Top-*k*, *k*=640

Pumping Station #3 shut down due to construction damage Find more at: www.abc.net.au/environment/species-worry/in-the-top-10-killer-whale-catastrophes-in-history.html
“In the top 10 killer whale catastrophes in history:
1) 1986: Up to 12 orcas struck by lightning; many drowned and many more badly injured.

Top-*k*, *k*=40, *t*=0.7

The whale's fate was confirmed late last week when the animal was found by fishermen off the coast of Bundaberg. Experts believe the whale was struck by a fishing vessel off the coast of Bundaberg, and died after being sucked into the ocean. The whale's fate was confirmed late last week when the animal was found by fishermen off the coast of Bundaberg.

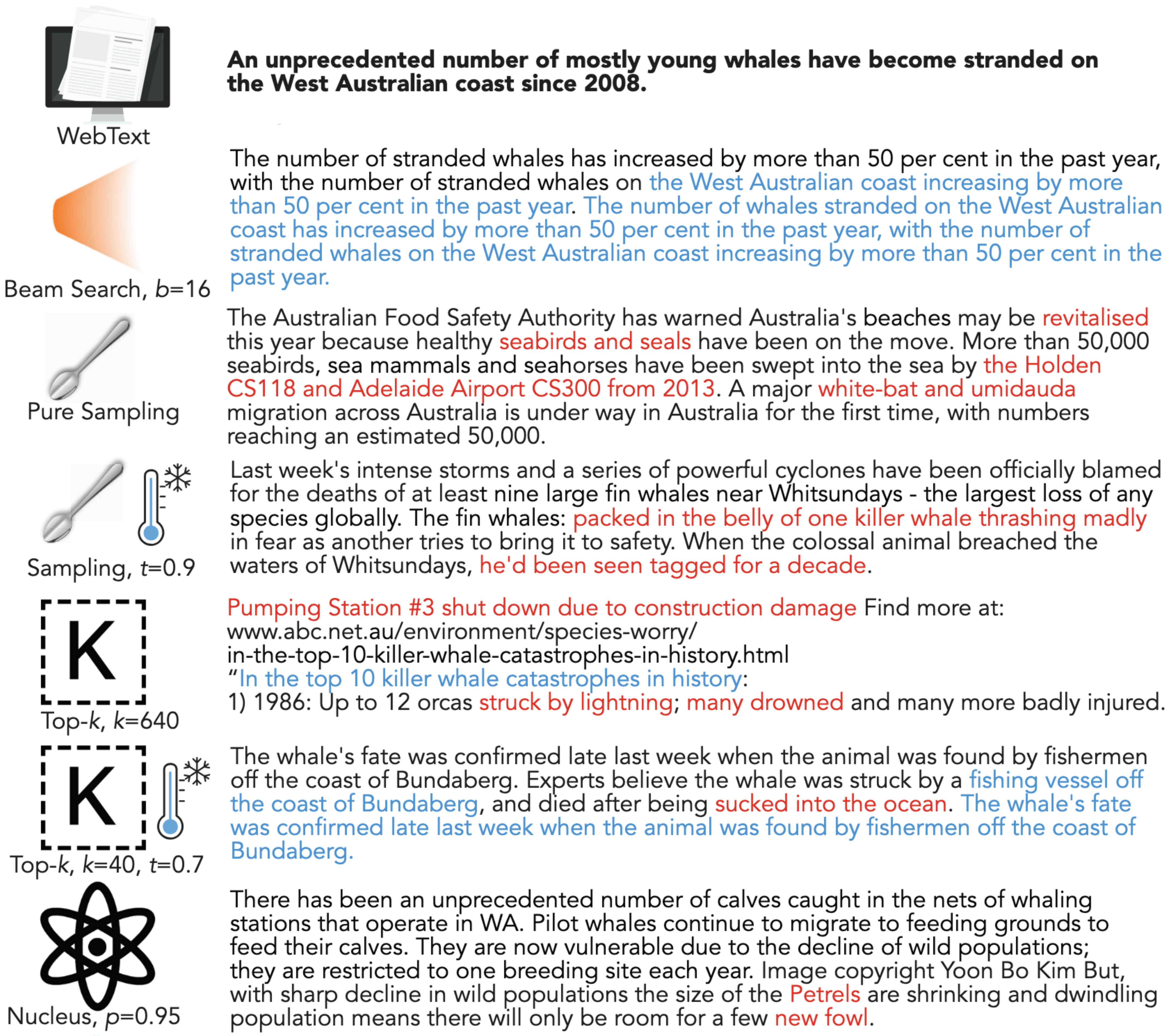
Nucleus, *p*=0.95

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.

Holtzman et al (2020)

Generation Big Picture

- In practice, we often **combine generation techniques**
  - E.g. top-k and temperature
- They have **differing drawbacks**
  - Greedy, Top-k, and Beam Search often lead to **repetitive** generation (blue)
  - Random sampling and high temperature lead to **nonsensical** generation (red)



WebText

An unprecedented number of mostly young whales have become stranded on the West Australian coast since 2008.

Beam Search, *b*=16

The number of stranded whales has increased by more than 50 per cent in the past year, with the number of stranded whales on the West Australian coast increasing by more than 50 per cent in the past year. The number of whales stranded on the West Australian coast has increased by more than 50 per cent in the past year, with the number of stranded whales on the West Australian coast increasing by more than 50 per cent in the past year.

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.

Top-*k*, *k*=640

Pumping Station #3 shut down due to construction damage Find more at: www.abc.net.au/environment/species-worry/in-the-top-10-killer-whale-catastrophes-in-history.html
“In the top 10 killer whale catastrophes in history:
1) 1986: Up to 12 orcas struck by lightning; many drowned and many more badly injured.

Top-*k*, *k*=40, *t*=0.7

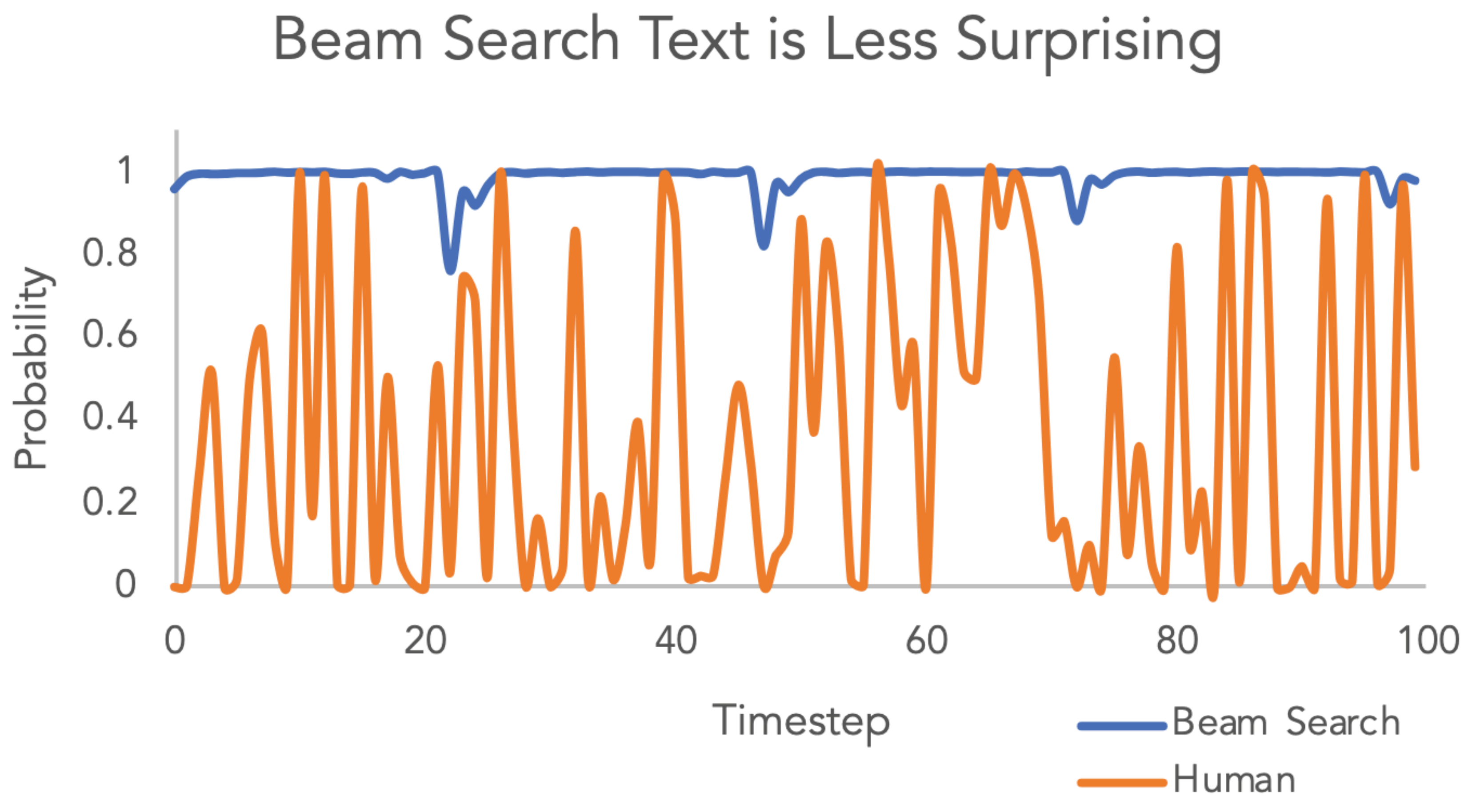
The whale's fate was confirmed late last week when the animal was found by fishermen off the coast of Bundaberg. Experts believe the whale was struck by a fishing vessel off the coast of Bundaberg, and died after being sucked into the ocean. The whale's fate was confirmed late last week when the animal was found by fishermen off the coast of Bundaberg.

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.

Holtzman et al (2020)

Probability and Information



Beam Search

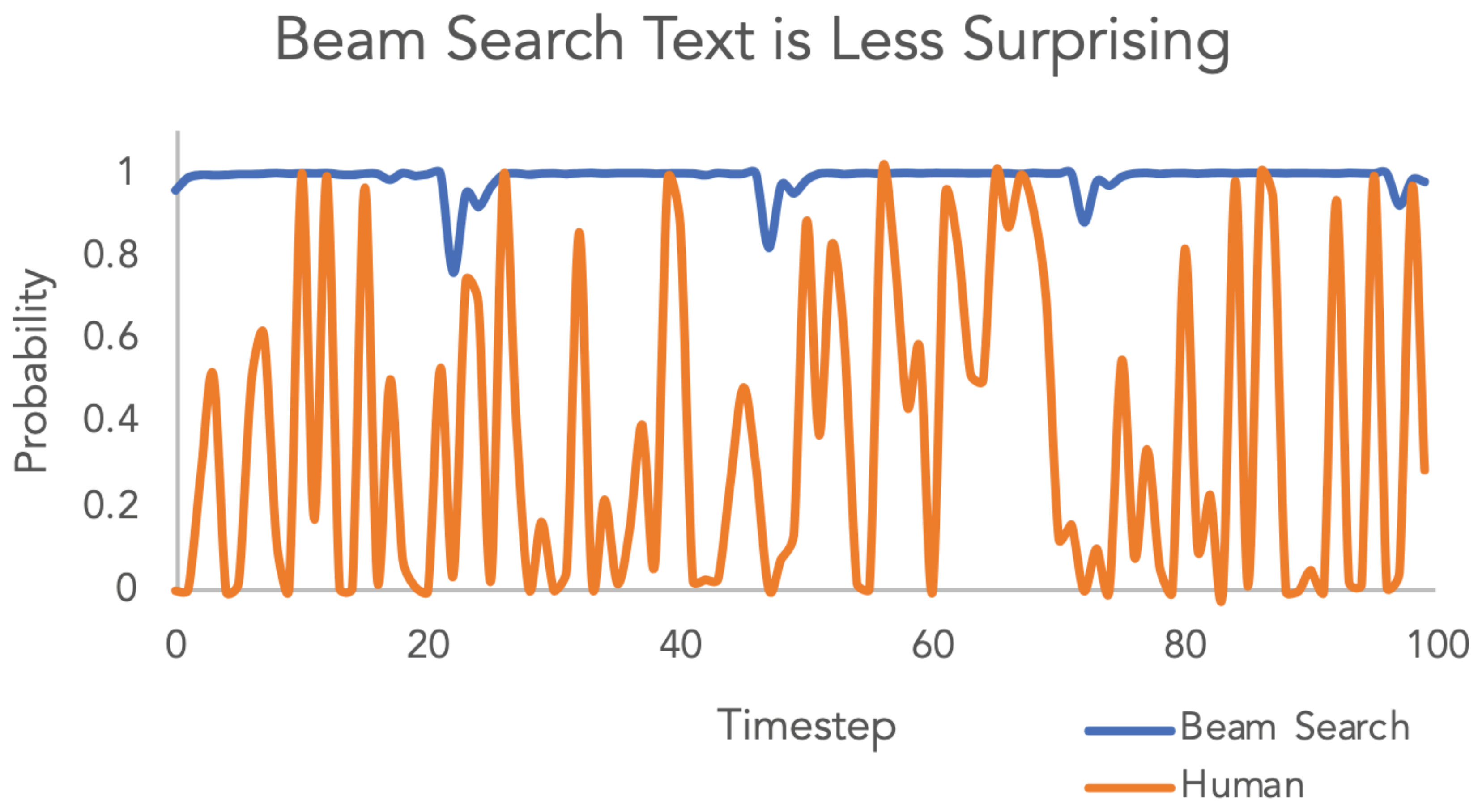
...to provide an overview of the current state-of-the-art in the field of computer vision and machine learning, and to provide an overview of the current state-of-the-art in the field of computer vision and machine learning, and to provide an overview of the current state-of-the-art in the field of computer vision and machine learning, and to provide an overview of the current state-of-the-art in the field of computer vision and machine learning, and...

Human

...which grant increased life span and three years warranty. The Antec HCG series consists of five models with capacities spanning from 400W to 900W. Here we should note that we have already tested the HCG-620 in a previous review and were quite satisfied With its performance. In today's review we will rigorously test the Antec HCG-520, which as its model number implies, has 520W capacity and contrary to Antec's strong beliefs in multi-rail PSUs is equipped...

Probability and Information

- Real human language **does not optimize for high probability!!!**



Beam Search

...to provide an overview of the current state-of-the-art in the field of computer vision and machine learning, and to provide an overview of the current state-of-the-art in the field of computer vision and machine learning, and to provide an overview of the current state-of-the-art in the field of computer vision and machine learning, and to provide an overview of the current state-of-the-art in the field of computer vision and machine learning, and...

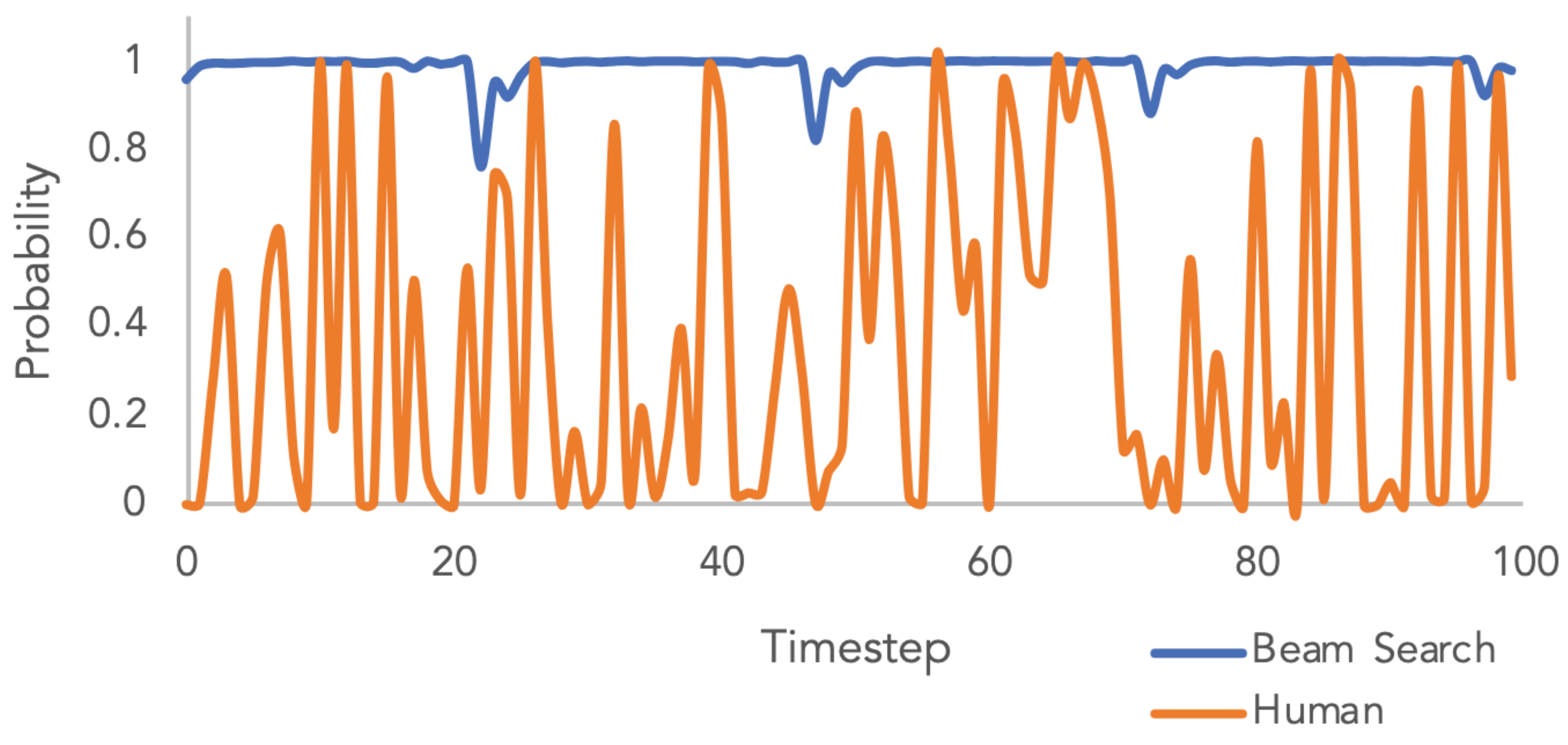
Human

...which grant increased life span and three years warranty. The Antec HCG series consists of five models with capacities spanning from 400W to 900W. Here we should note that we have already tested the HCG-620 in a previous review and were quite satisfied With its performance. In today's review we will rigorously test the Antec HCG-520, which as its model number implies, has 520W capacity and contrary to Antec's strong beliefs in multi-rail PSUs is equipped...

Probability and Information

- Real human language **does not optimize for high probability!!!**
- Actually, in **Information Theory**, **low probability → high information!**

Beam Search Text is Less Surprising



Beam Search

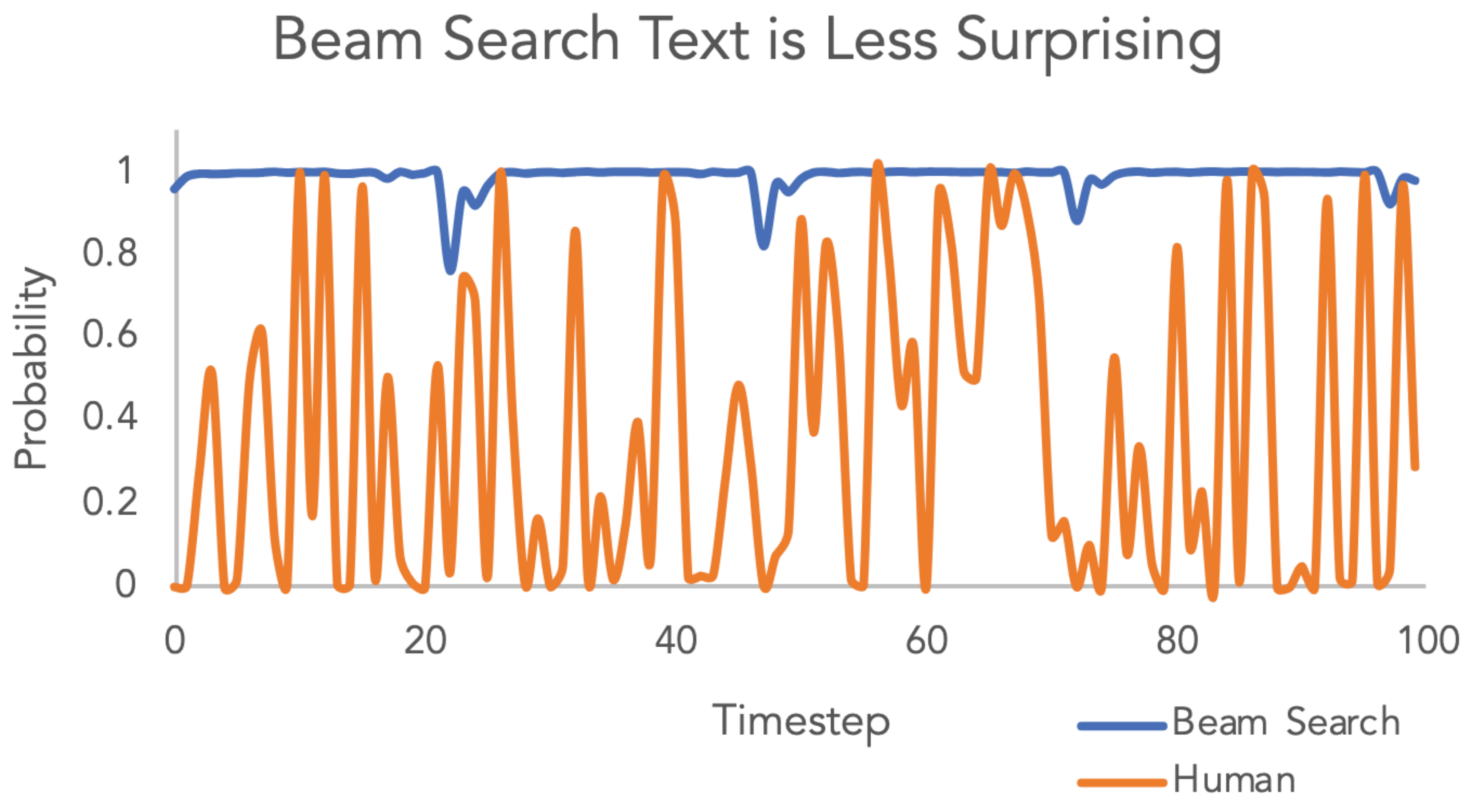
...to provide an overview of the current state-of-the-art in the field of computer vision and machine learning, and to provide an overview of the current state-of-the-art in the field of computer vision and machine learning, and to provide an overview of the current state-of-the-art in the field of computer vision and machine learning, and to provide an overview of the current state-of-the-art in the field of computer vision and machine learning, and...

Human

...which grant increased life span and three years warranty. The Antec HCG series consists of five models with capacities spanning from 400W to 900W. Here we should note that we have already tested the HCG-620 in a previous review and were quite satisfied With its performance. In today's review we will rigorously test the Antec HCG-520, which as its model number implies, has 520W capacity and contrary to Antec's strong beliefs in multi-rail PSUs is equipped...

Probability and Information

- Real human language **<u>does not optimize for high probability!!!</u>**
- Actually, in **Information Theory**, **low probability → high information!**
    - In order to convey information, humans **must use low-probability symbols**



Beam Search

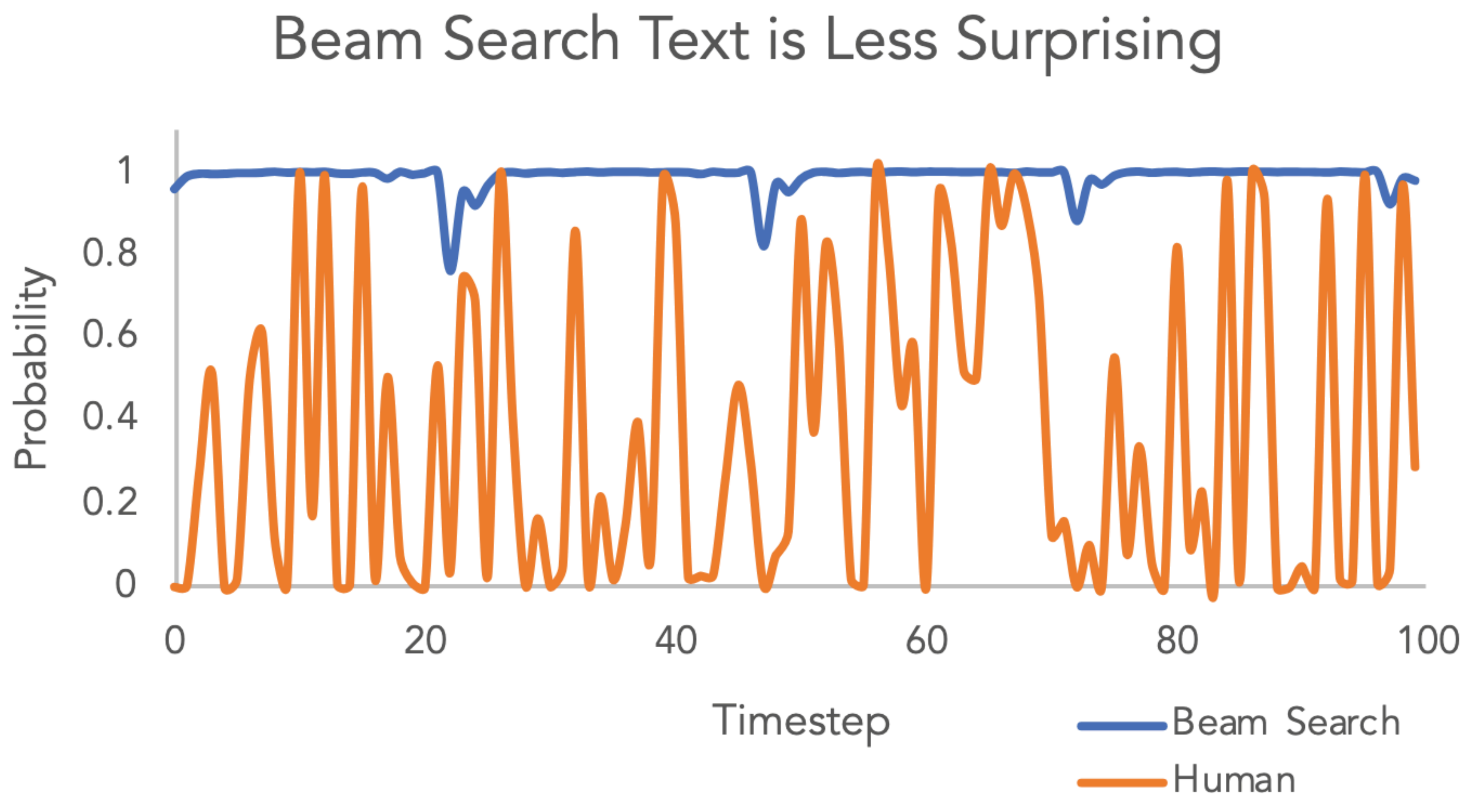
...to provide an overview of the current state-of-the-art in the field of computer vision and machine learning, and to provide an overview of the current state-of-the-art in the field of computer vision and machine learning, and to provide an overview of the current state-of-the-art in the field of computer vision and machine learning, and to provide an overview of the current state-of-the-art in the field of computer vision and machine learning, and...

Human

...which grant increased life span and three years warranty. The Antec HCG series consists of five models with capacities spanning from 400W to 900W. Here we should note that we have already tested the HCG-620 in a previous review and were quite satisfied With its performance. In today's review we will rigorously test the Antec HCG-520, which as its model number implies, has 520W capacity and contrary to Antec's strong beliefs in multi-rail PSUs is equipped...

Probability and Information

- Real human language **does not optimize for high probability!!!**
- Actually, in **Information Theory**, **low probability → high information!**
  - In order to convey information, humans **must use low-probability symbols**
- Techniques like Beam Search **don't emulate** human language well in this regard



Beam Search

...to provide an overview of the current state-of-the-art in the field of computer vision and machine learning, and to provide an overview of the current state-of-the-art in the field of computer vision and machine learning, and to provide an overview of the current state-of-the-art in the field of computer vision and machine learning, and to provide an overview of the current state-of-the-art in the field of computer vision and machine learning, and...

Human

...which grant increased life span and three years warranty. The Antec HCG series consists of five models with capacities spanning from 400W to 900W. Here we should note that we have already tested the HCG-620 in a previous review and were quite satisfied With its performance. In today's review we will rigorously test the Antec HCG-520, which as its model number implies, has 520W capacity and contrary to Antec's strong beliefs in multi-rail PSUs is equipped...

Important Points

Important Points

- Generation needs a **stopping condition** (it's a **while-loop**!)

Important Points

- Generation needs a **stopping condition** (it's a **while-loop**!)
    - Common: **End-Of-Sequence** symbol (</s>, <eos>) or **maximum length**

Important Points

- Generation needs a **stopping condition** (it's a **while-loop**!)
  - Common: **End-Of-Sequence** symbol (</s>, <eos>) or **maximum length**
- We usually have a **trade-off** between **probability and "creativity"**

Important Points

- Generation needs a **stopping condition** (it's a **while-loop**!)
    - Common: **End-Of-Sequence** symbol (</s>, <eos>) or **maximum length**
- We usually have a **trade-off** between **probability and "creativity"**
- Modern **Chatbots / "Large Language Models"** are trained for **more than just Language Modeling** (i.e. next word prediction)

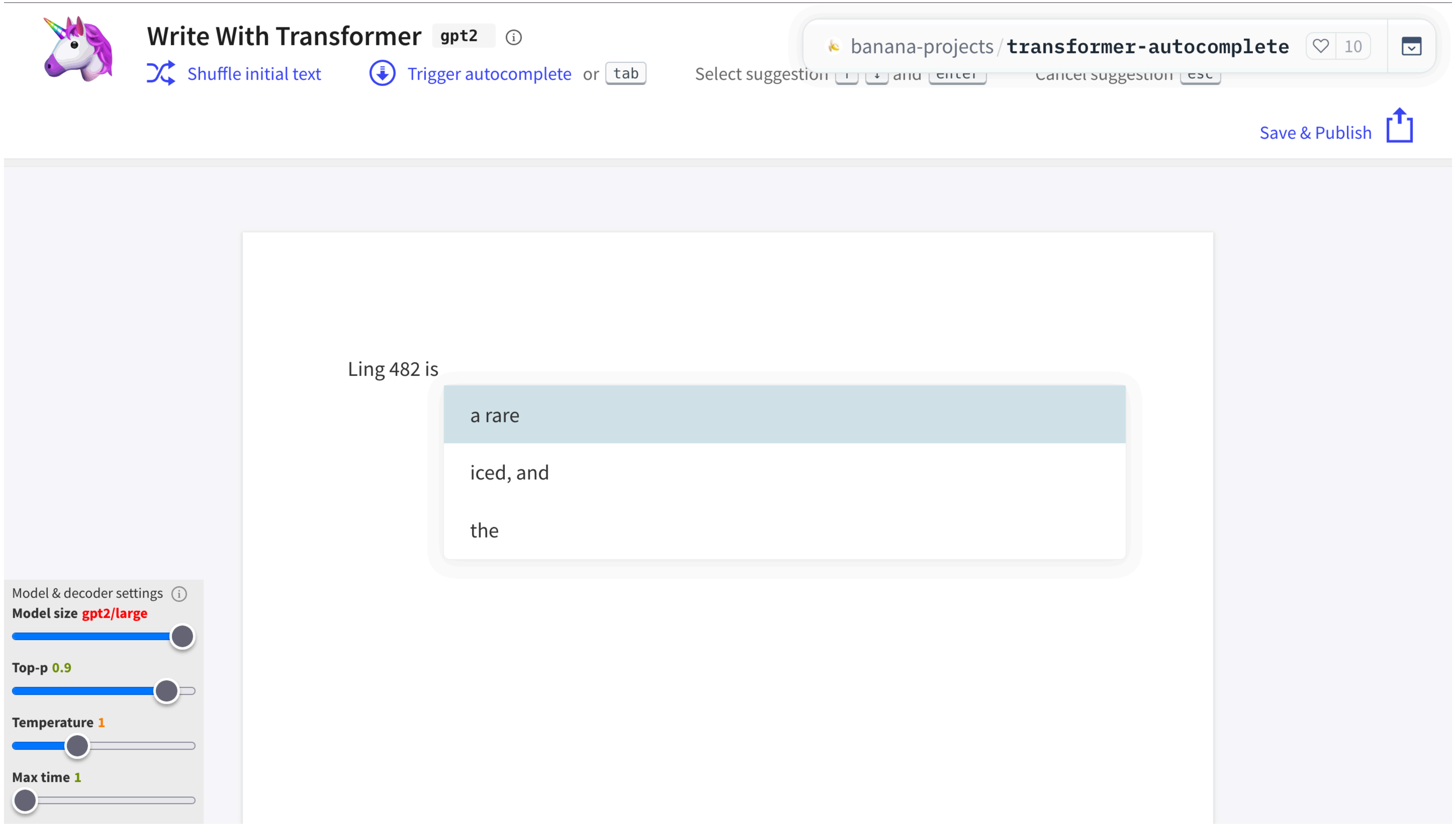
Important Points

- Generation needs a **stopping condition** (it's a **while-loop**!)
    - Common: **End-Of-Sequence** symbol (</s>, <eos>) or **maximum length**
- We usually have a **trade-off** between **probability and "creativity"**
- Modern **Chatbots / "Large Language Models"** are trained for **more than just Language Modeling** (i.e. next word prediction)
    - This **drastically changes the way they generate**

Important Points

- Generation needs a **stopping condition** (it's a **while-loop**!)
    - Common: **End-Of-Sequence** symbol (</s>, <eos>) or **maximum length**
- We usually have a **trade-off** between **probability and "creativity"**
- Modern **Chatbots / "Large Language Models"** are trained for **more than just Language Modeling** (i.e. next word prediction)
    - This **drastically changes the way they generate**
    - "Pure" Language Models try to complete a prompt. Chatbots try to respond to a prompt (more later)

Write With Transformer



Write With Transformer gpt2
Shuffle initial text
Trigger autocomplete or tab
Select suggestion ↑ ↓ and enter
Cancel suggestion esc
banana-projects / transformer-autocomplete
10
Save & Publish

Ling 482 is

a rare

iced, and

the

Model & decoder settings

Model size gpt2/large

Top-p 0.9

Temperature 1

Max time 1