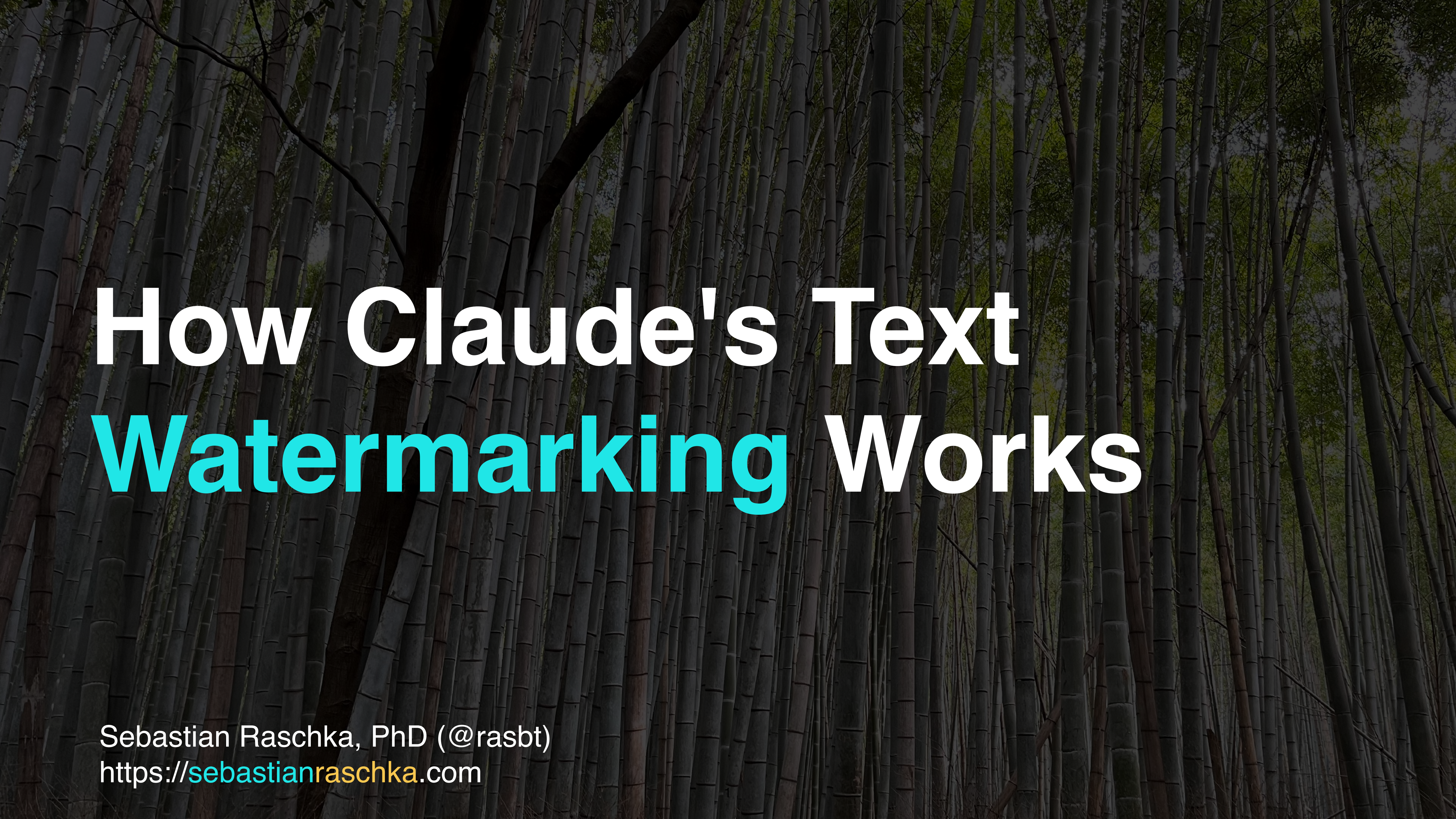




How Claude's Text Watermarking Works

Sebastian Raschka, PhD (@rasbt)
<https://sebastianraschka.com>

Benefits of “from scratch”

(and how conventional LLMs work...)

anthropic.com

How Claude's text watermarking works | Anthropic

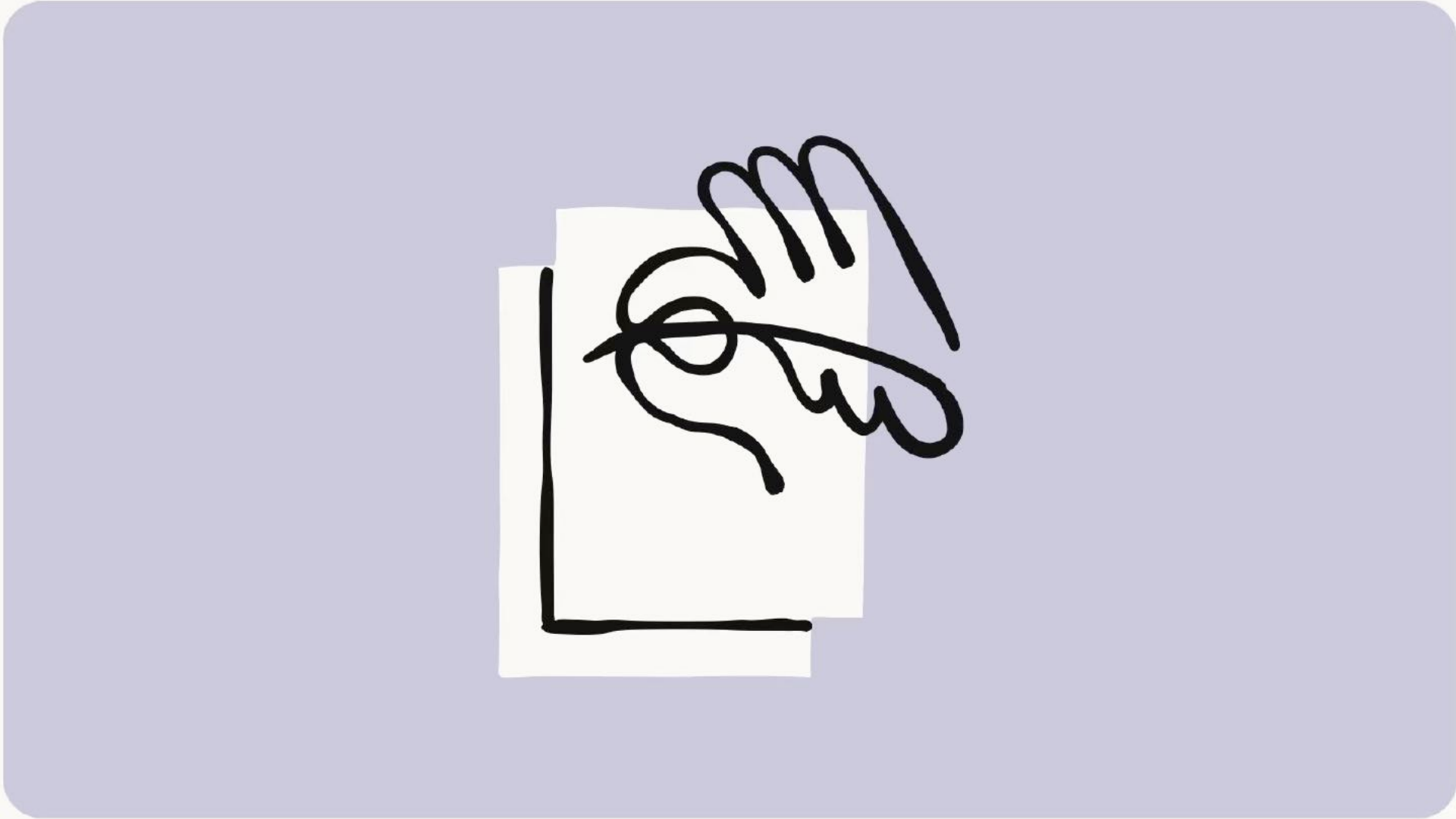
ANTHROPIC

ResearchPolicyCommitmentsLearnNewsTry Claude

Announcements

How Claude's text watermark works

Aug 14, 2026



What is watermarking?

How does watermarking affect Claude's outputs?

Future Claude models will generate text that contains a watermark. This is a way of determining the likelihood that Claude was involved in writing the text, and we, along with several other major AI providers, are implementing this change to comply with the EU AI Act.

<https://www.anthropic.com/news/claude-text-watermark>

Prelude: text generation in LLMs

Chat

Work

How can I help, Seb?

+ Ask ChatGPT

High ▾

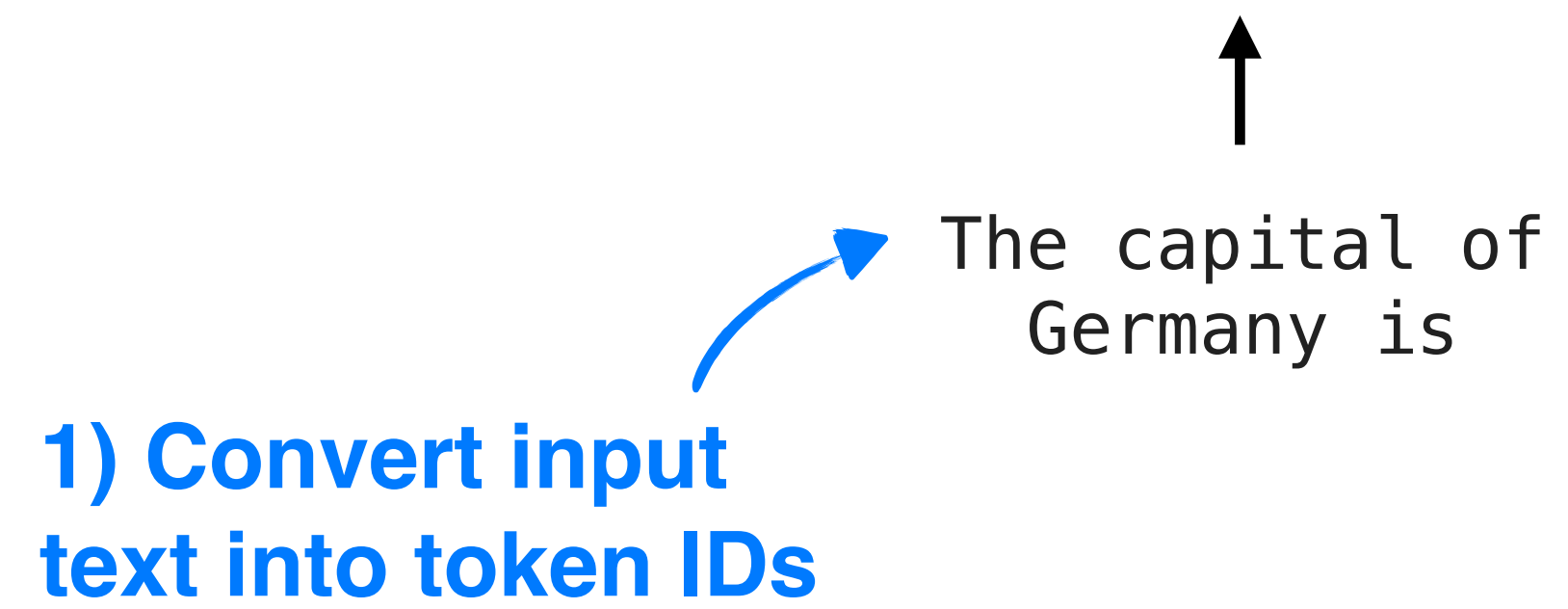


Review the Jekyll conversion status and next actions each week

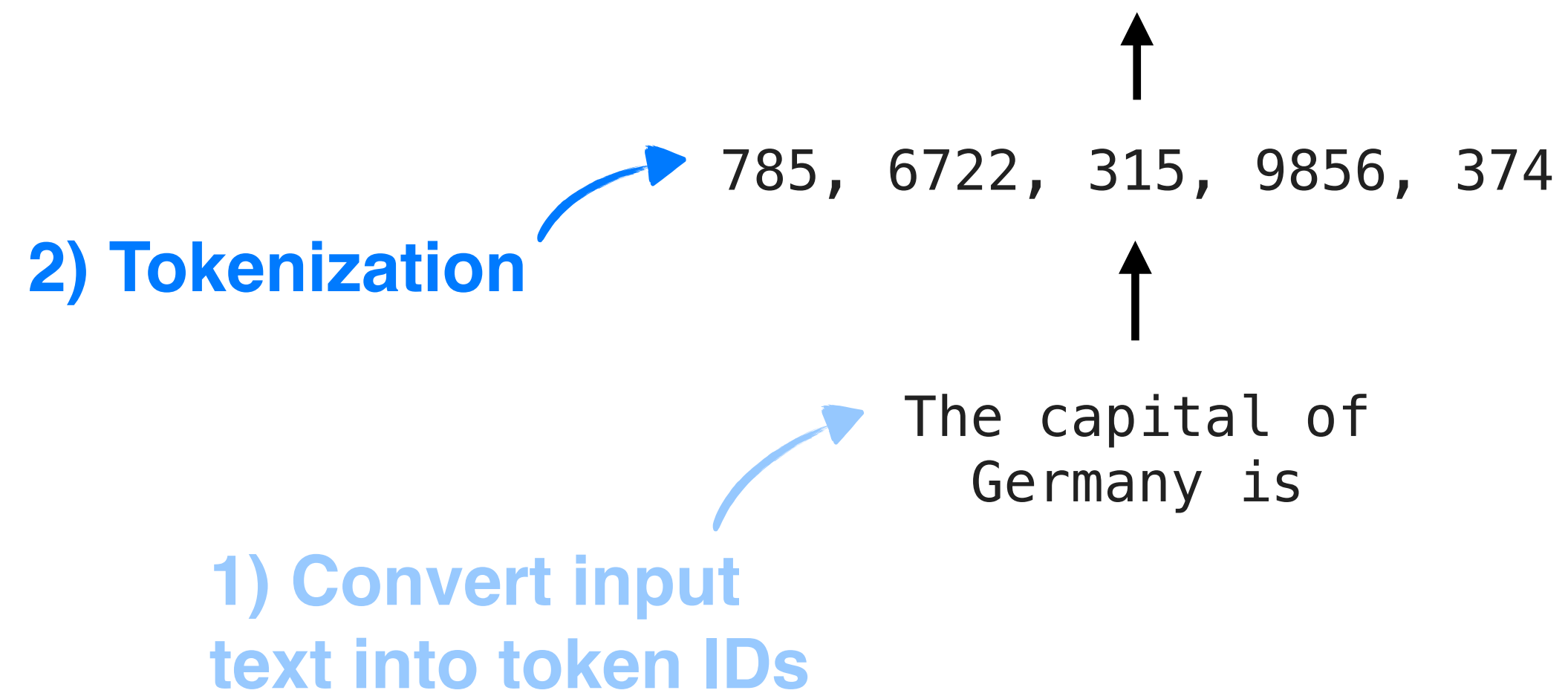


**So what happens under the hood?
when the **next token** is generated?**

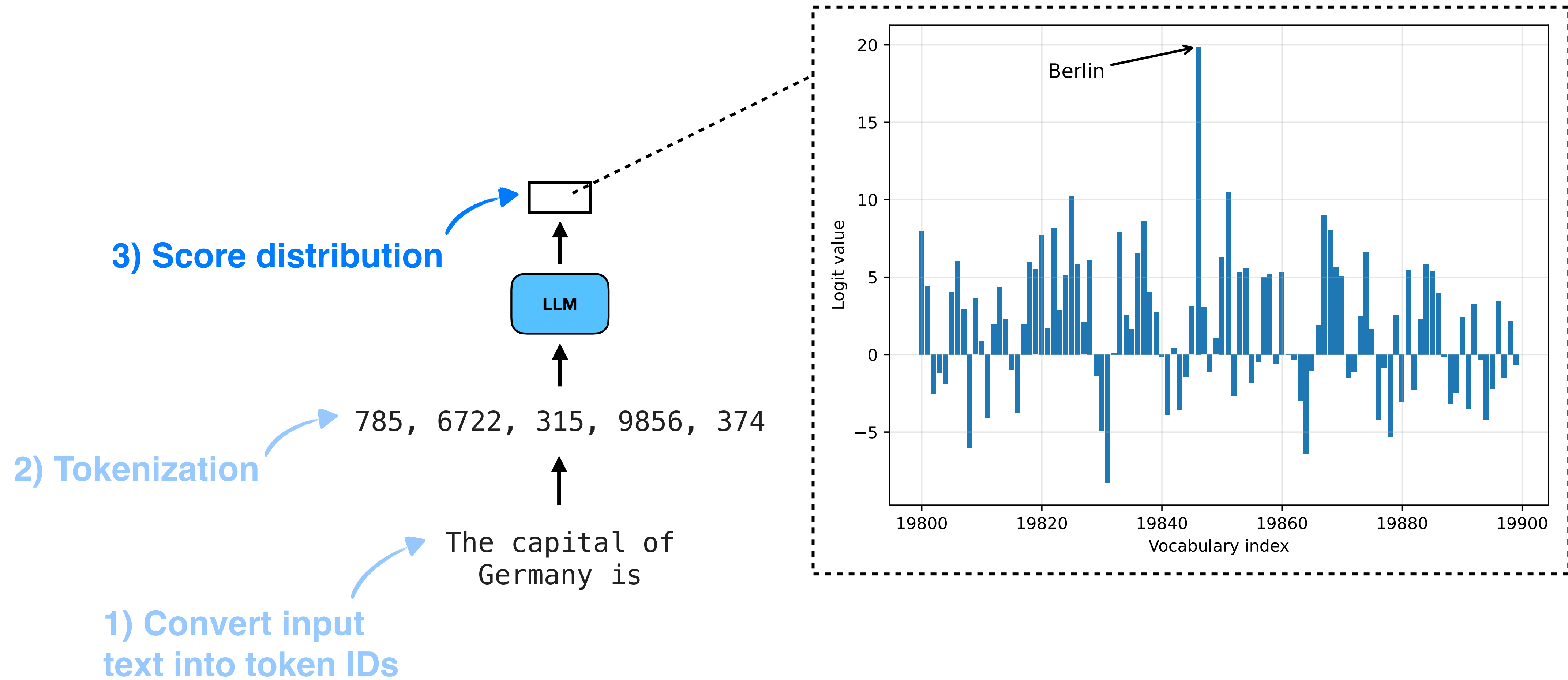
Generating the next token



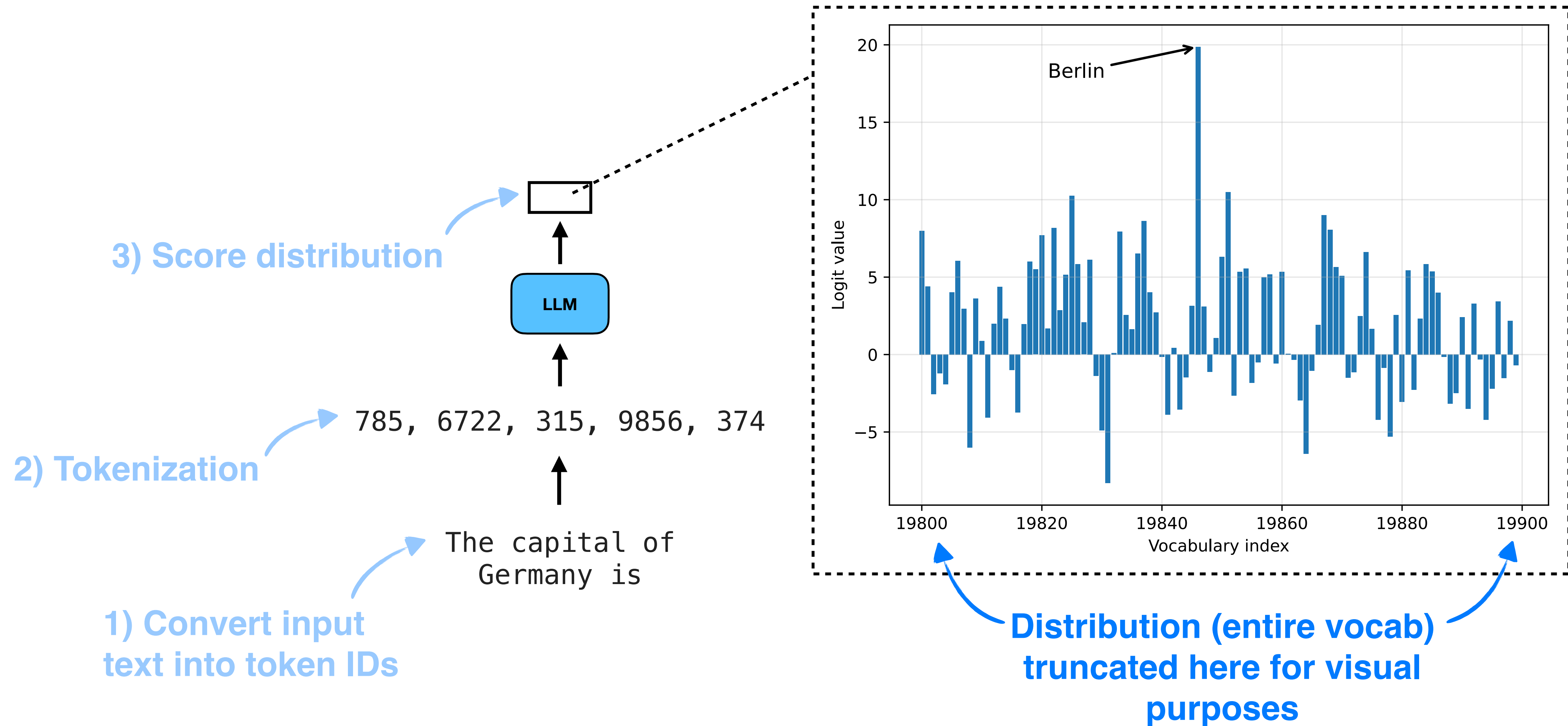
Generating the next token



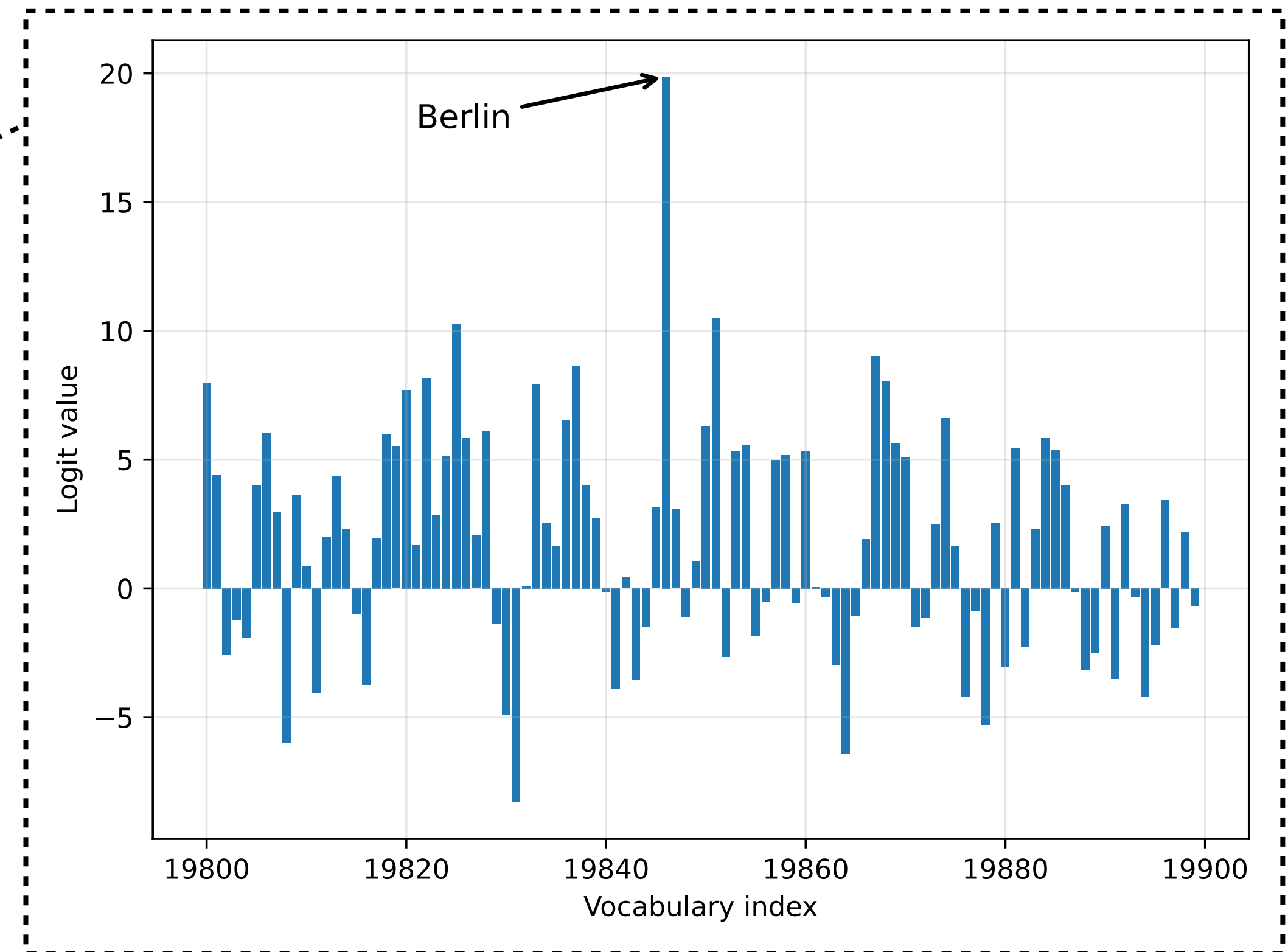
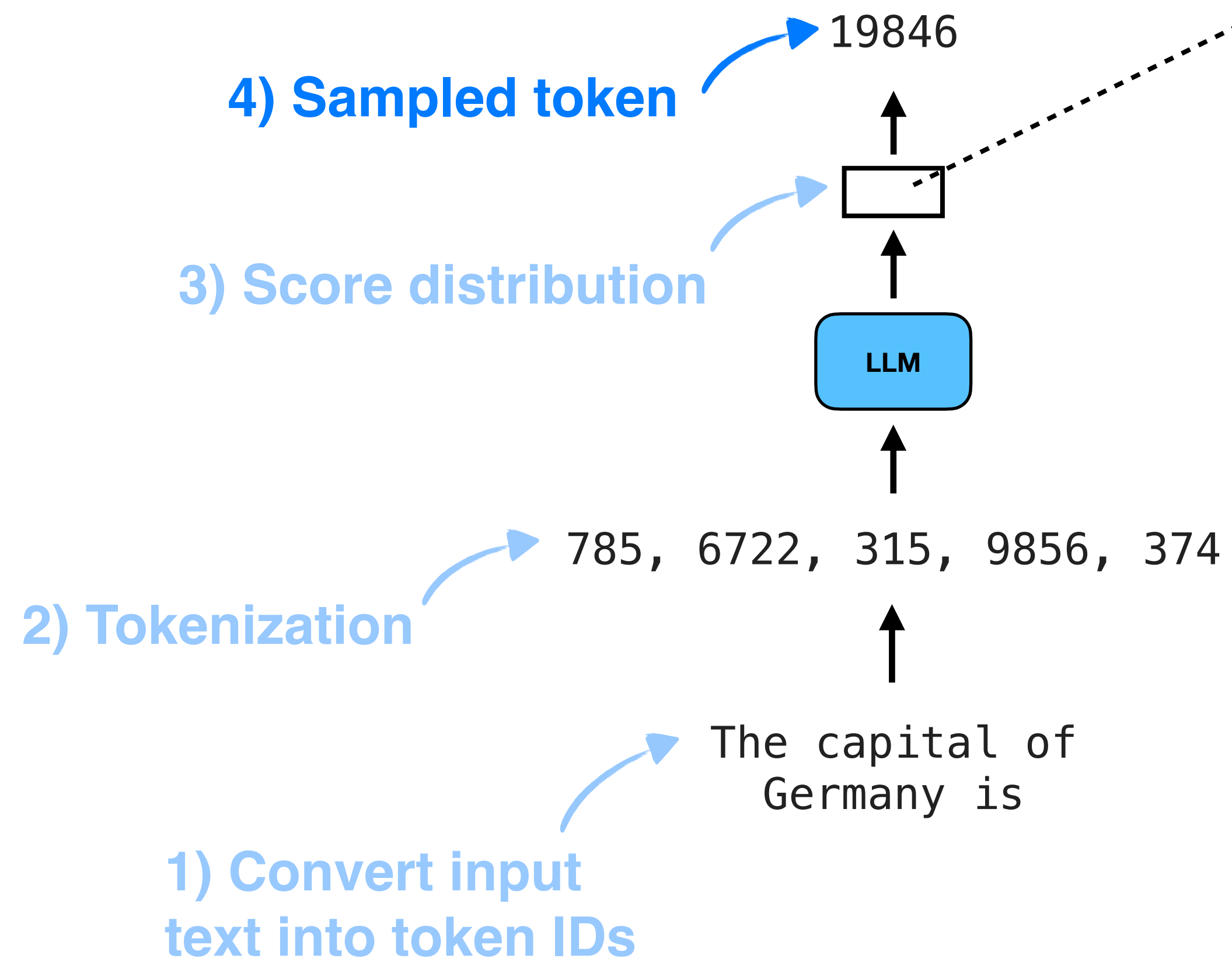
Generating the next token



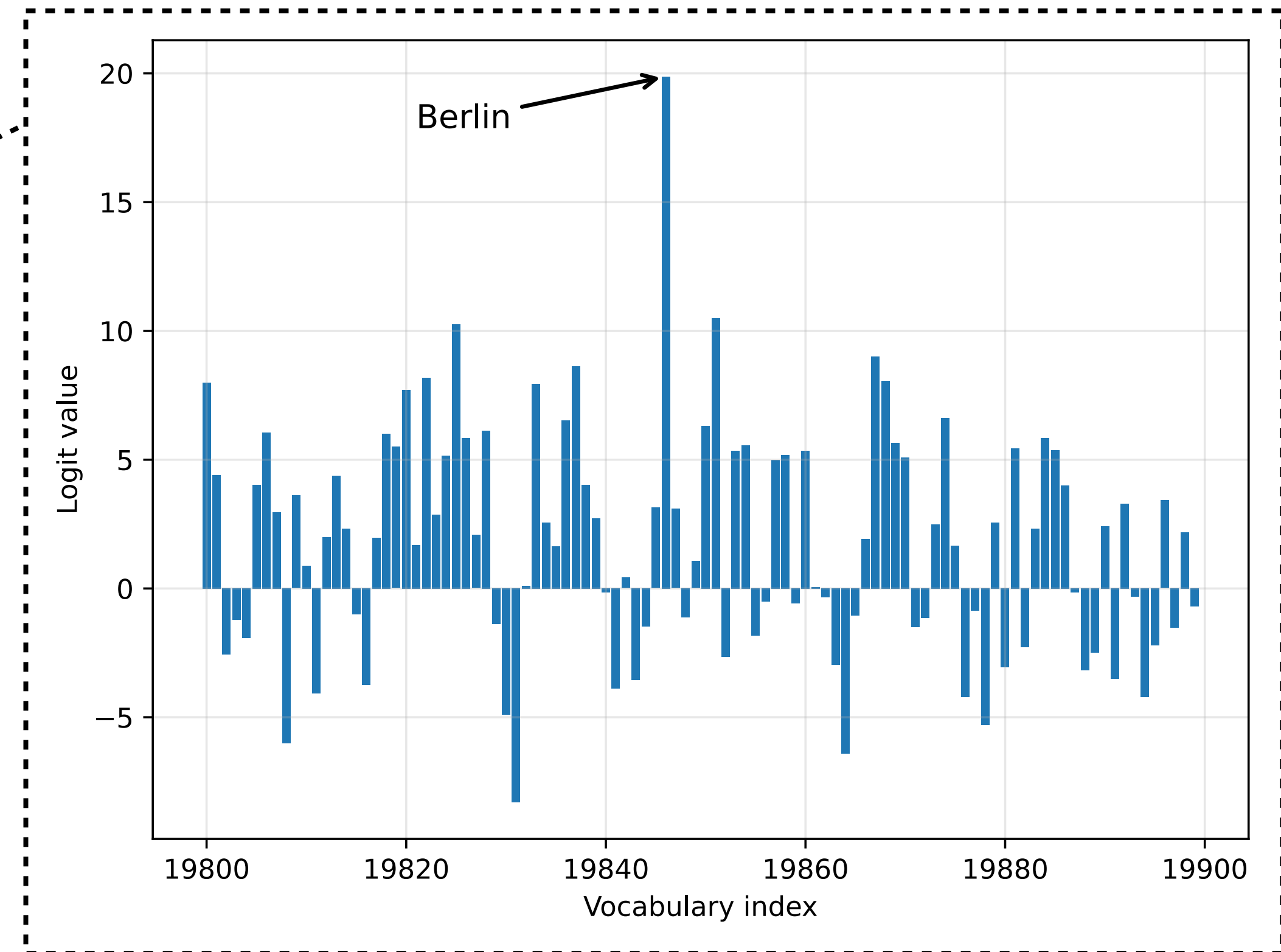
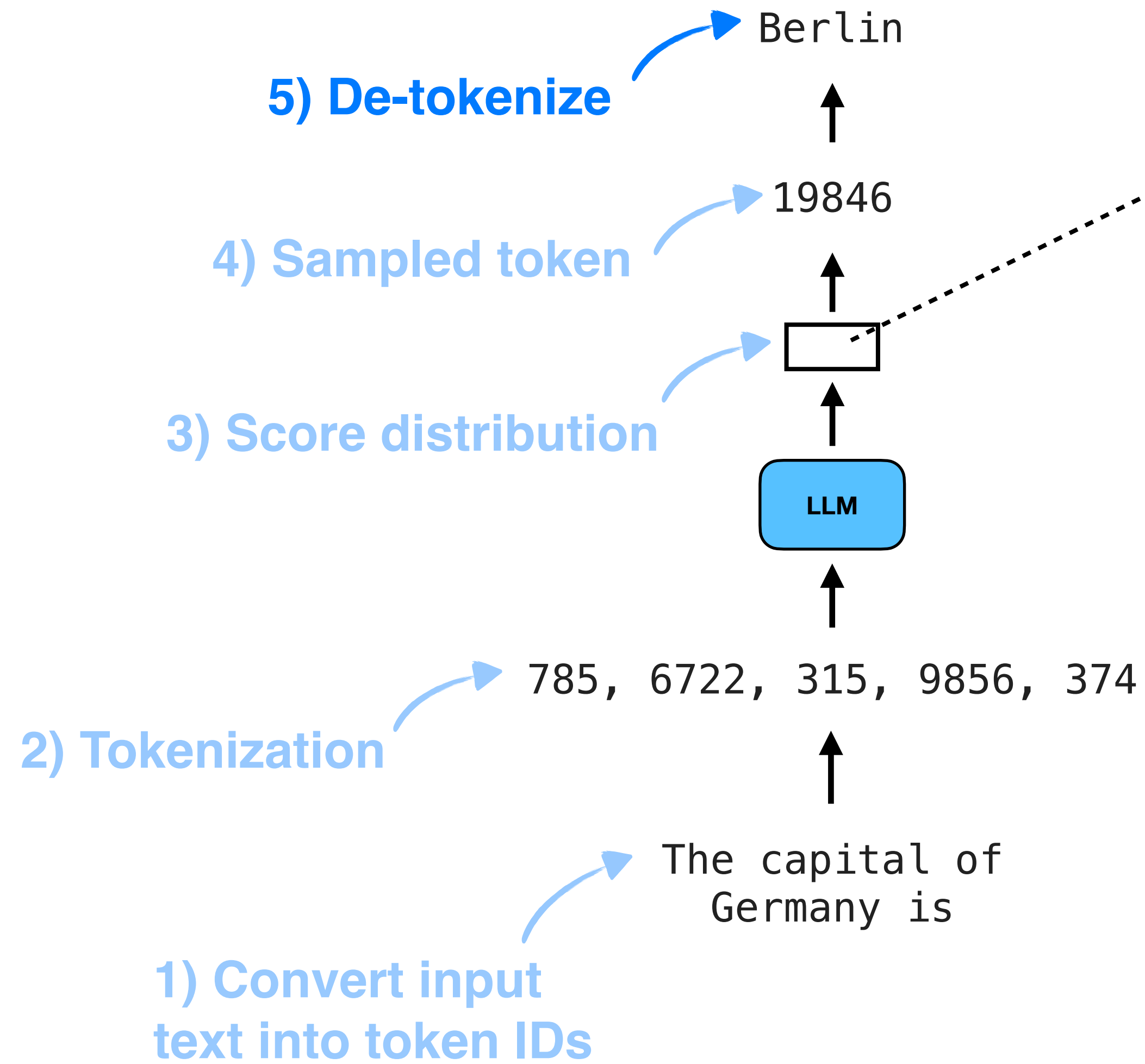
Generating the next token



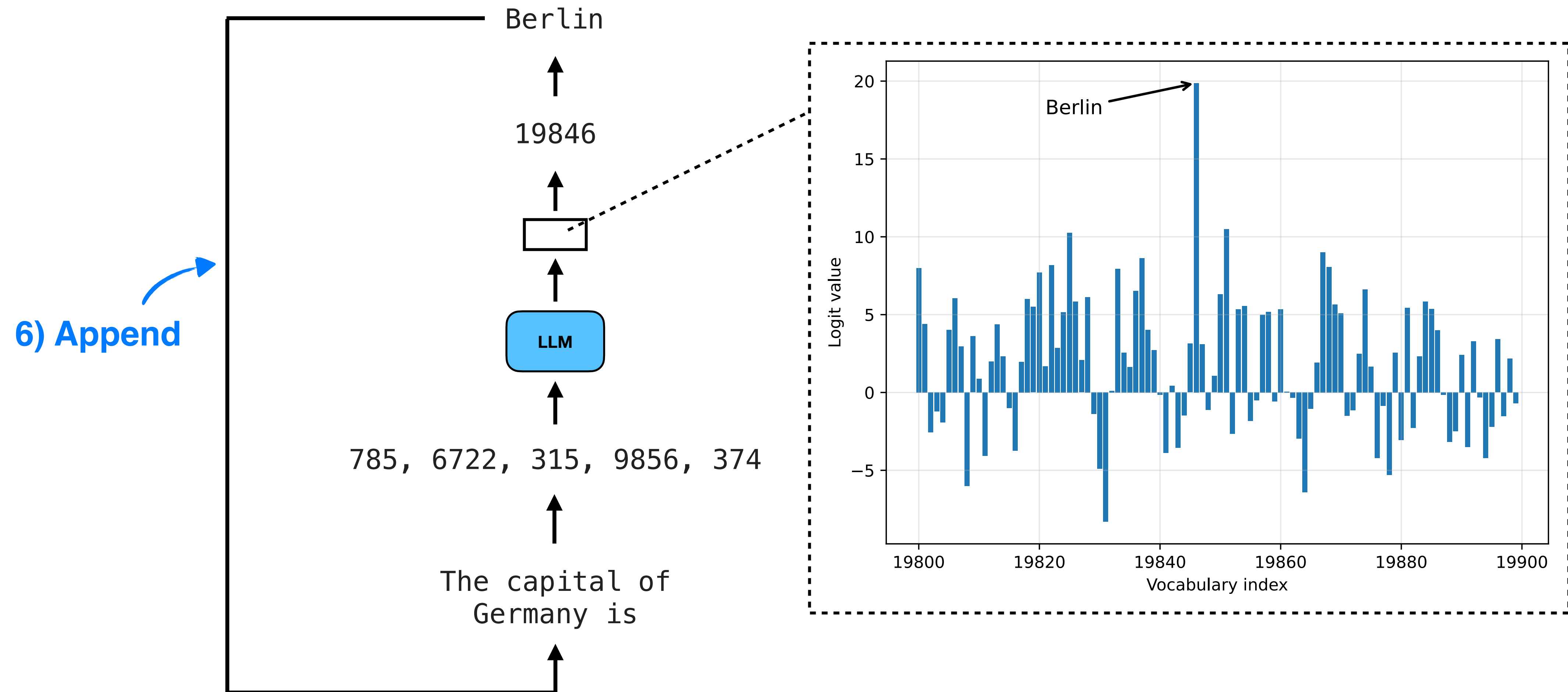
Generating the next token



Generating the next token

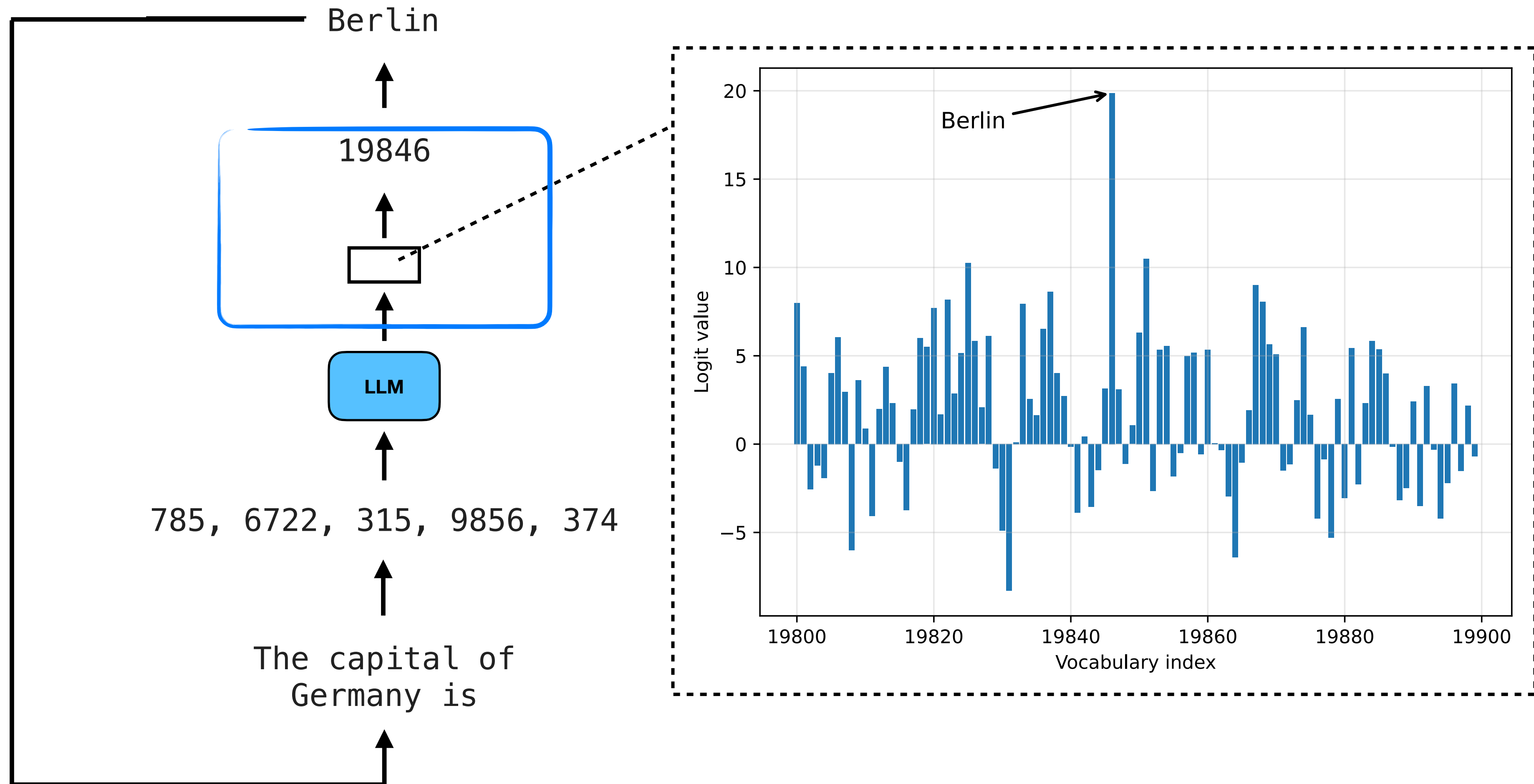


Generating the next token

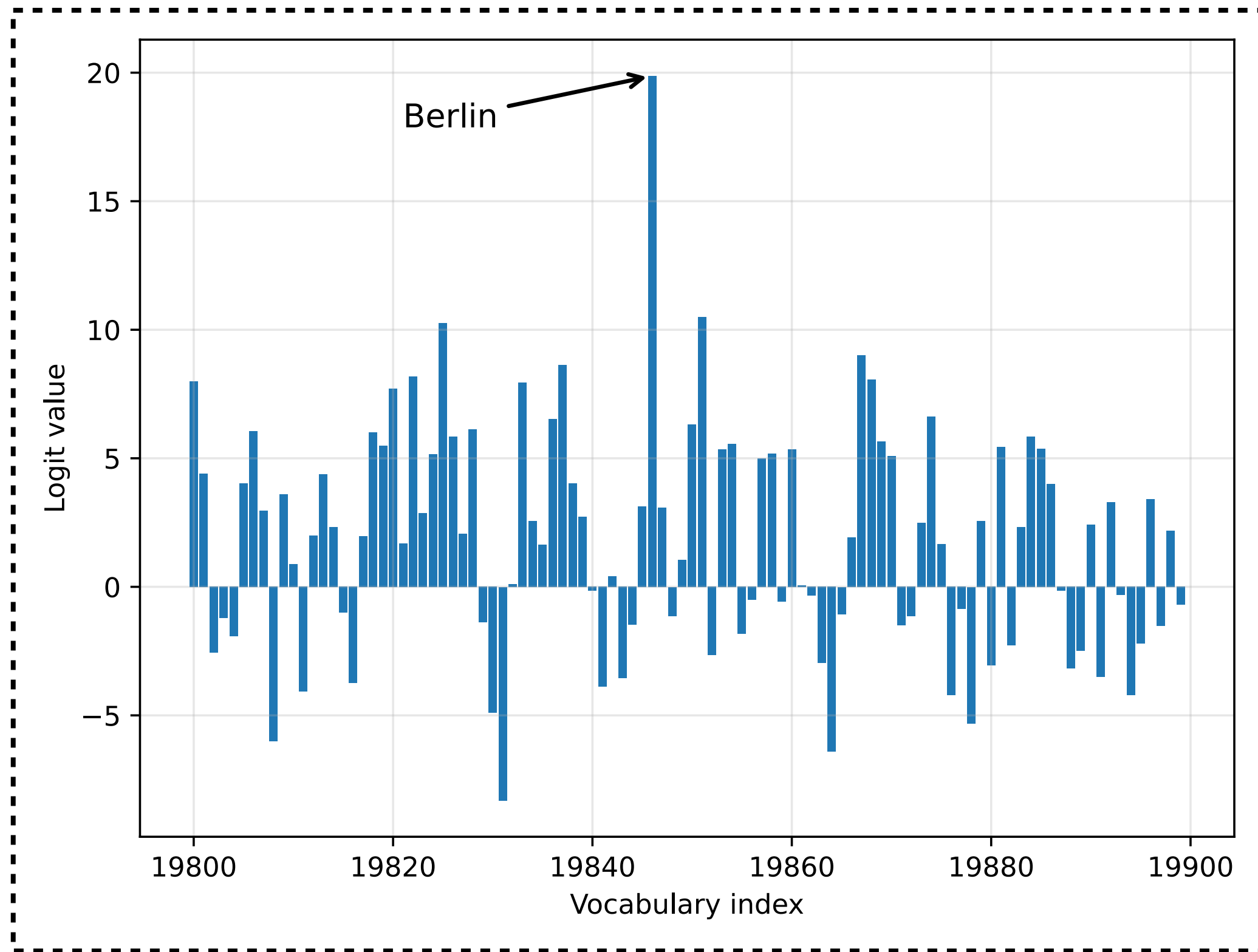


Generating the next token

How do we
sample the
next token?



Token sampling



```
import numpy as np

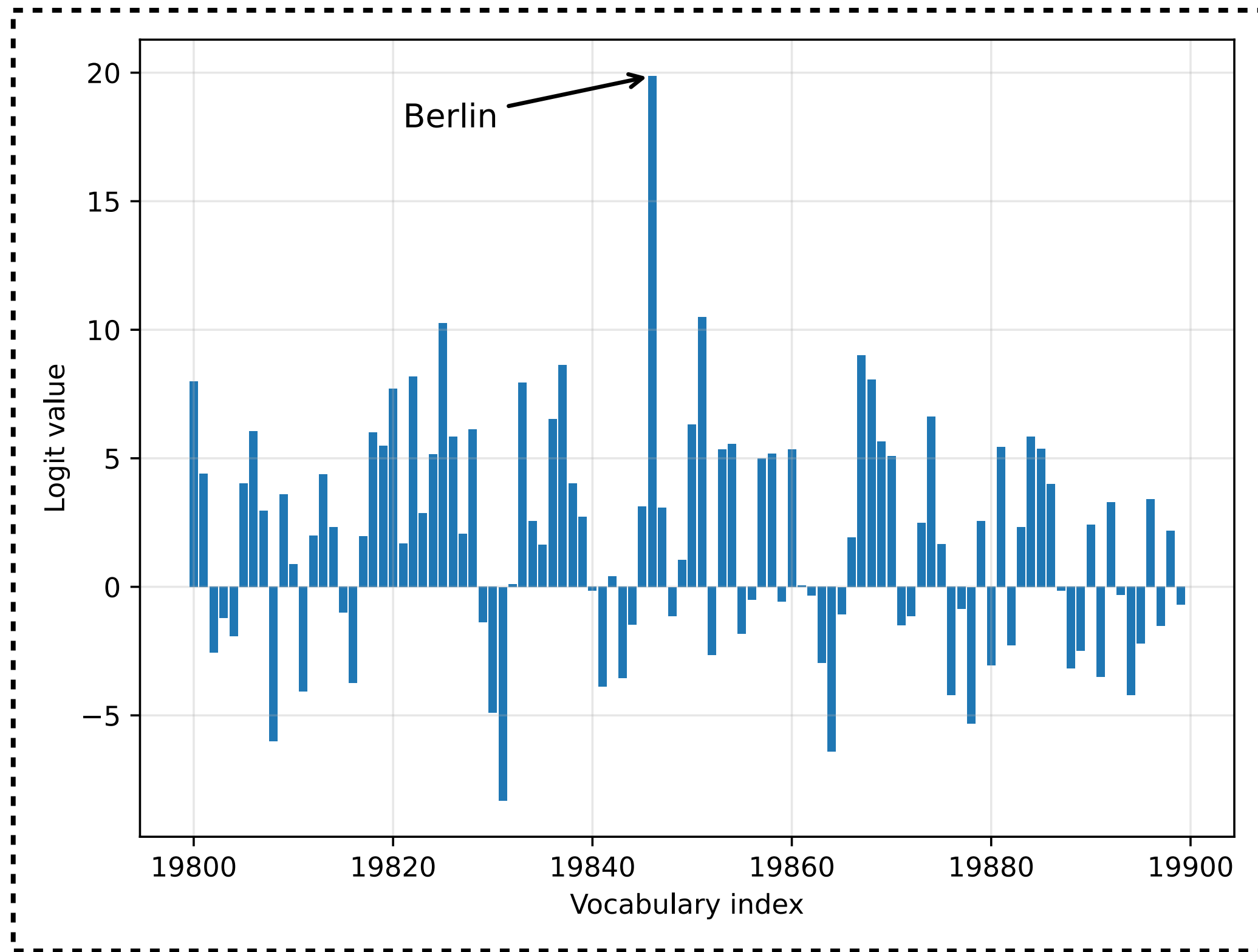
scores = np.array([
    7.9,  4.4, -2.6, -1.2, -1.9,  4.0,  6.0,  3.0, -6.0,  3.6,
    0.8, -4.0,  2.0,  4.4,  2.3, -1.0, -3.7,  1.9,  6.0,  5.5,
    7.7,  1.7,  8.2,  2.8,  5.1, 10.2,  5.8,  2.1,  6.1, -1.3,
   -4.9, -8.3,  0.1,  7.9,  2.6,  1.7,  6.5,  8.6,  4.0,  2.7,
   -0.2, -3.9,  0.4, -3.6, -1.5,  3.1, 19.8,  3.1, -1.1,  1.0,
    6.3, 10.5, -2.6,  5.3,  5.5, -1.8, -0.5,  5.0,  5.1, -0.6,
    5.3,  0.0, -0.3, -3.0, -6.4, -1.1,  1.9,  9.0,  8.0,  5.6,
    5.0, -1.5, -1.1,  2.5,  6.6,  1.7, -4.2, -0.8, -5.3,  2.6,
   -3.0,  5.4, -2.2,  2.3,  5.8,  5.4,  4.0, -0.2, -3.1, -2.5,
    2.4, -3.5,  3.3, -0.3, -4.2, -2.1,  3.4, -1.5,  2.1, -0.7
])

vocab_indices = np.arange(19800, 19900)

# Softmax
probs = np.exp(scores) / np.exp(scores).sum()
```

Convert to probabilities (all values sum to 1)

Token sampling



```
import numpy as np

scores = np.array([
    7.9,  4.4, -2.6, -1.2, -1.9,  4.0,  6.0,  3.0, -6.0,  3.6,
    0.8, -4.0,  2.0,  4.4,  2.3, -1.0, -3.7,  1.9,  6.0,  5.5,
    7.7,  1.7,  8.2,  2.8,  5.1, 10.2,  5.8,  2.1,  6.1, -1.3,
   -4.9, -8.3,  0.1,  7.9,  2.6,  1.7,  6.5,  8.6,  4.0,  2.7,
   -0.2, -3.9,  0.4, -3.6, -1.5,  3.1, 19.8,  3.1, -1.1,  1.0,
    6.3, 10.5, -2.6,  5.3,  5.5, -1.8, -0.5,  5.0,  5.1, -0.6,
    5.3,  0.0, -0.3, -3.0, -6.4, -1.1,  1.9,  9.0,  8.0,  5.6,
    5.0, -1.5, -1.1,  2.5,  6.6,  1.7, -4.2, -0.8, -5.3,  2.6,
   -3.0,  5.4, -2.2,  2.3,  5.8,  5.4,  4.0, -0.2, -3.1, -2.5,
    2.4, -3.5,  3.3, -0.3, -4.2, -2.1,  3.4, -1.5,  2.1, -0.7
])

vocab_indices = np.arange(19800, 19900)

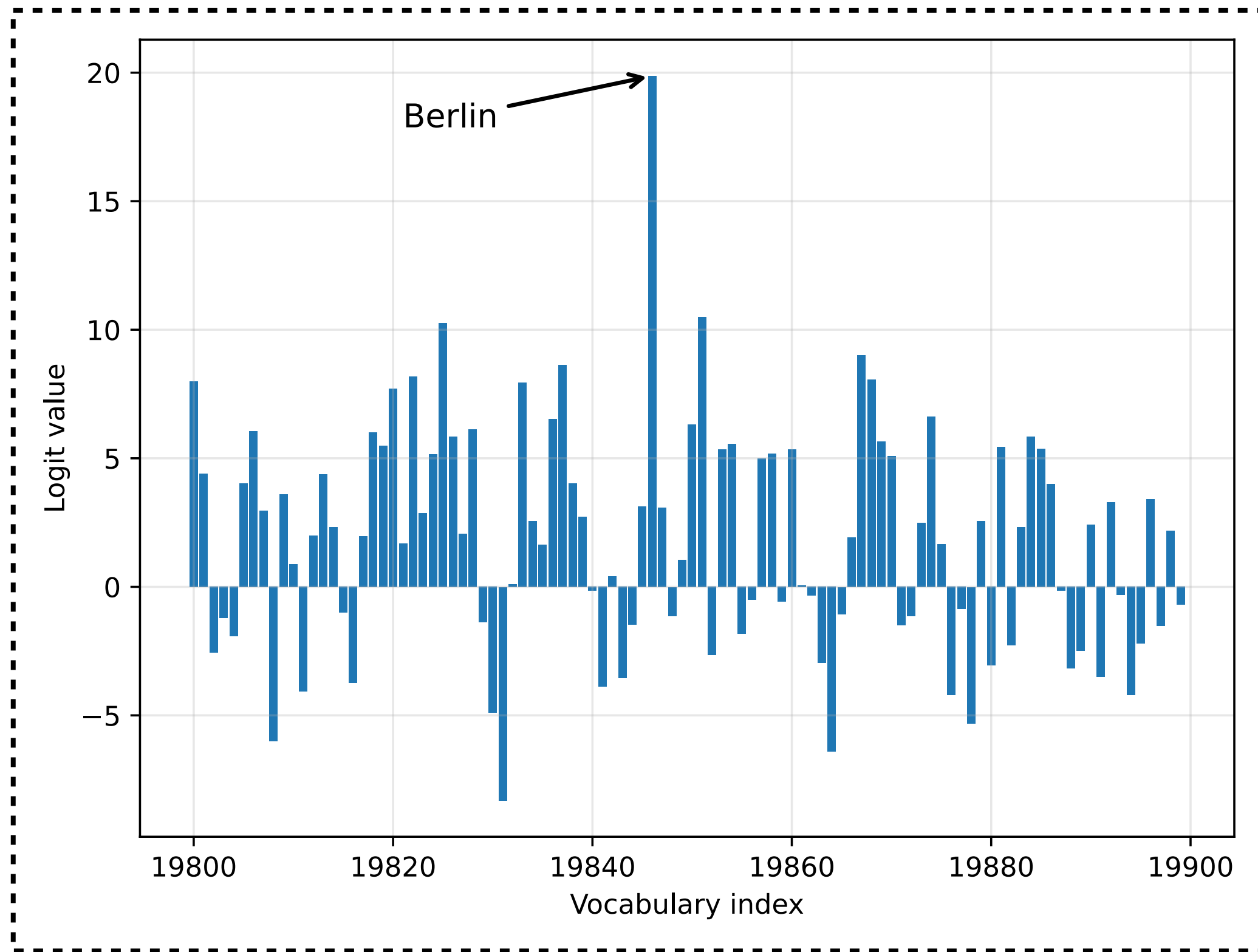
# Softmax
probs = np.exp(scores) / np.exp(scores).sum()

rng = np.random.default_rng(123)
sample = rng.choice(vocab_indices, p=probs)
print(f"{sample} -> {tokenizer.decode([sample])}")
```

19846 -> Berlin

**Sample token from probability distribution
(probability = “contribution weights”)**

Token sampling



```
samples = rng.choice(vocab_indices, size=10_000, replace=True, p=probs)
tokens, counts = np.unique(samples, return_counts=True)

top5 = np.argsort(counts)[-5:][::-1]

for sample, count in zip(tokens[top5], counts[top5]):
    print(f"{sample} -> {tokenizer.decode([sample]): {count}x")
```

19846 -> Berlin: 9997x
19851 -> Hal: 2x
19822 -> Moh: 1x

E.g., if we sample 10k times, “Berlin” would be selected most of the time

Announcements

How Claude's text watermark works

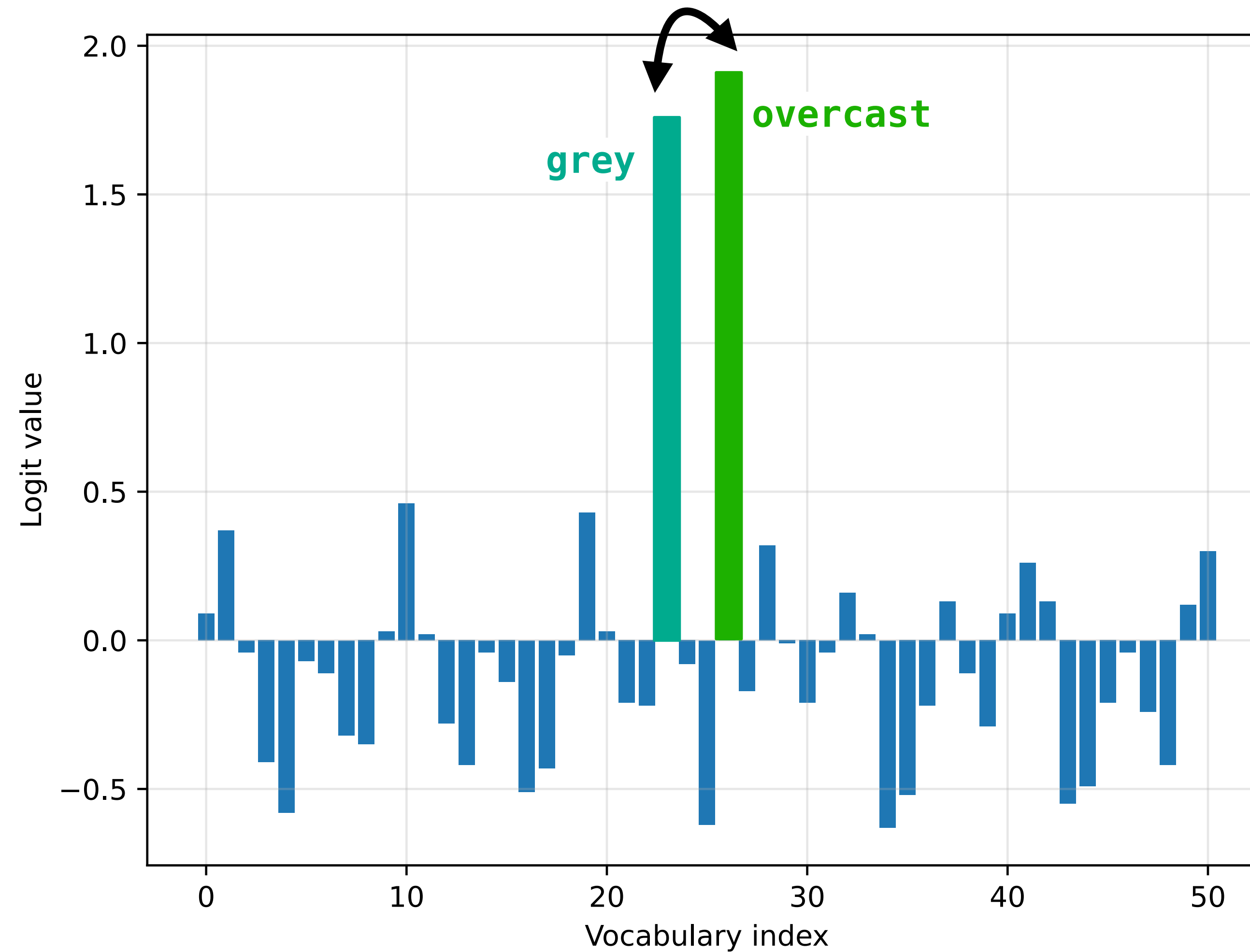
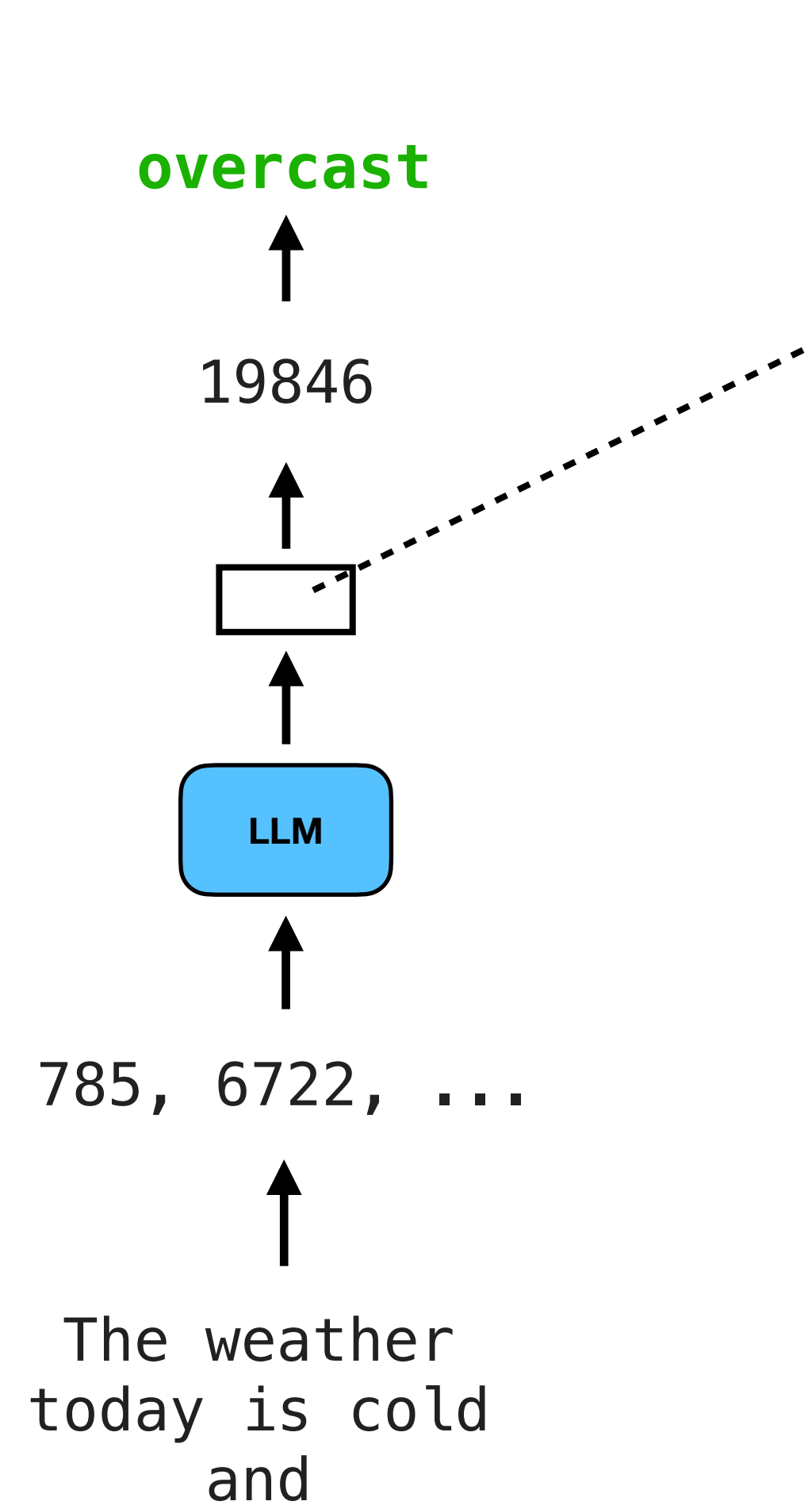
Aug 14, 2026



<https://www.anthropic.com/news/claude-text-watermark>

Without watermarking:

Due to random sampling we select either of these likely tokens



Random seeds add **reproducibility**

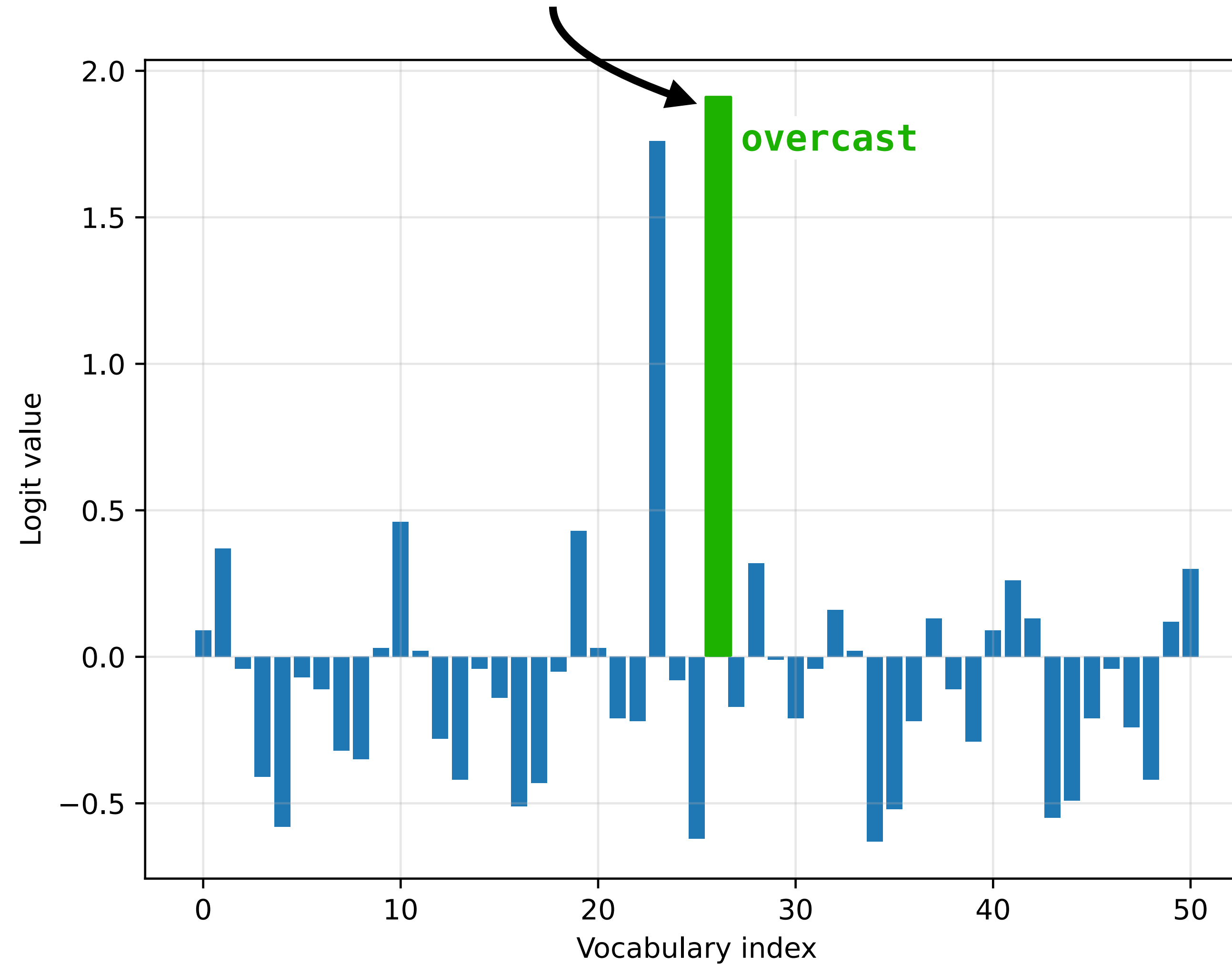
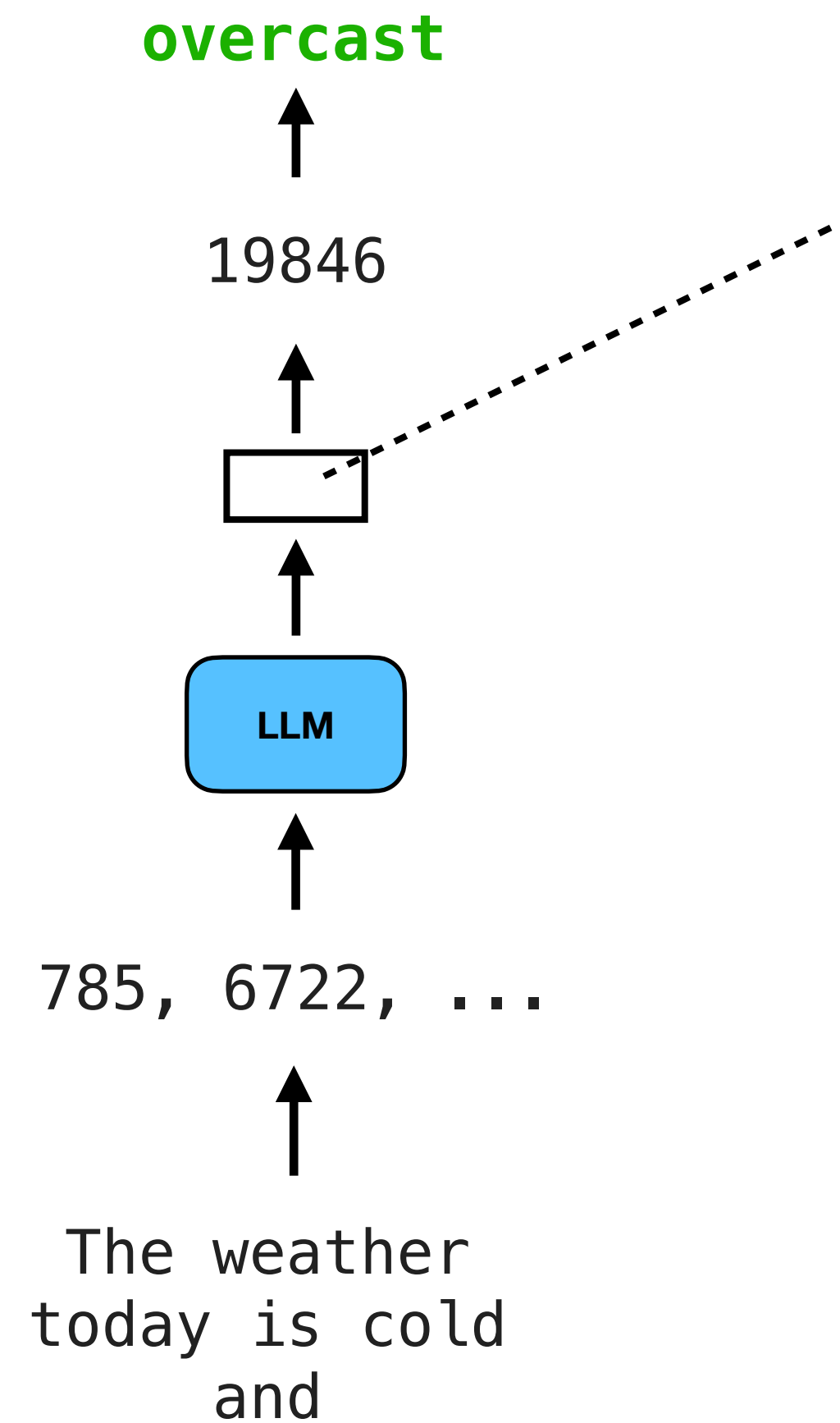
```
[ ]: import numpy as np
```



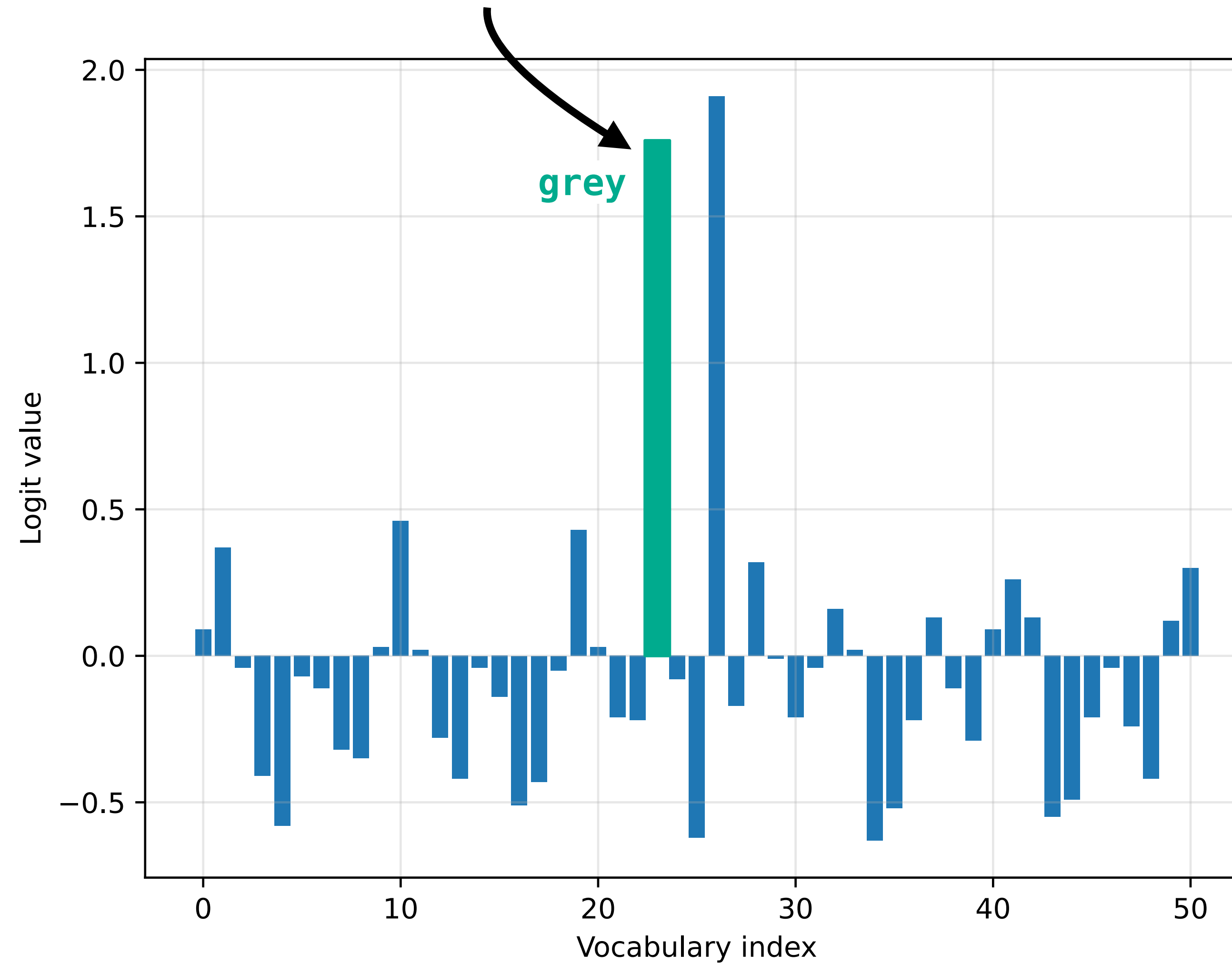
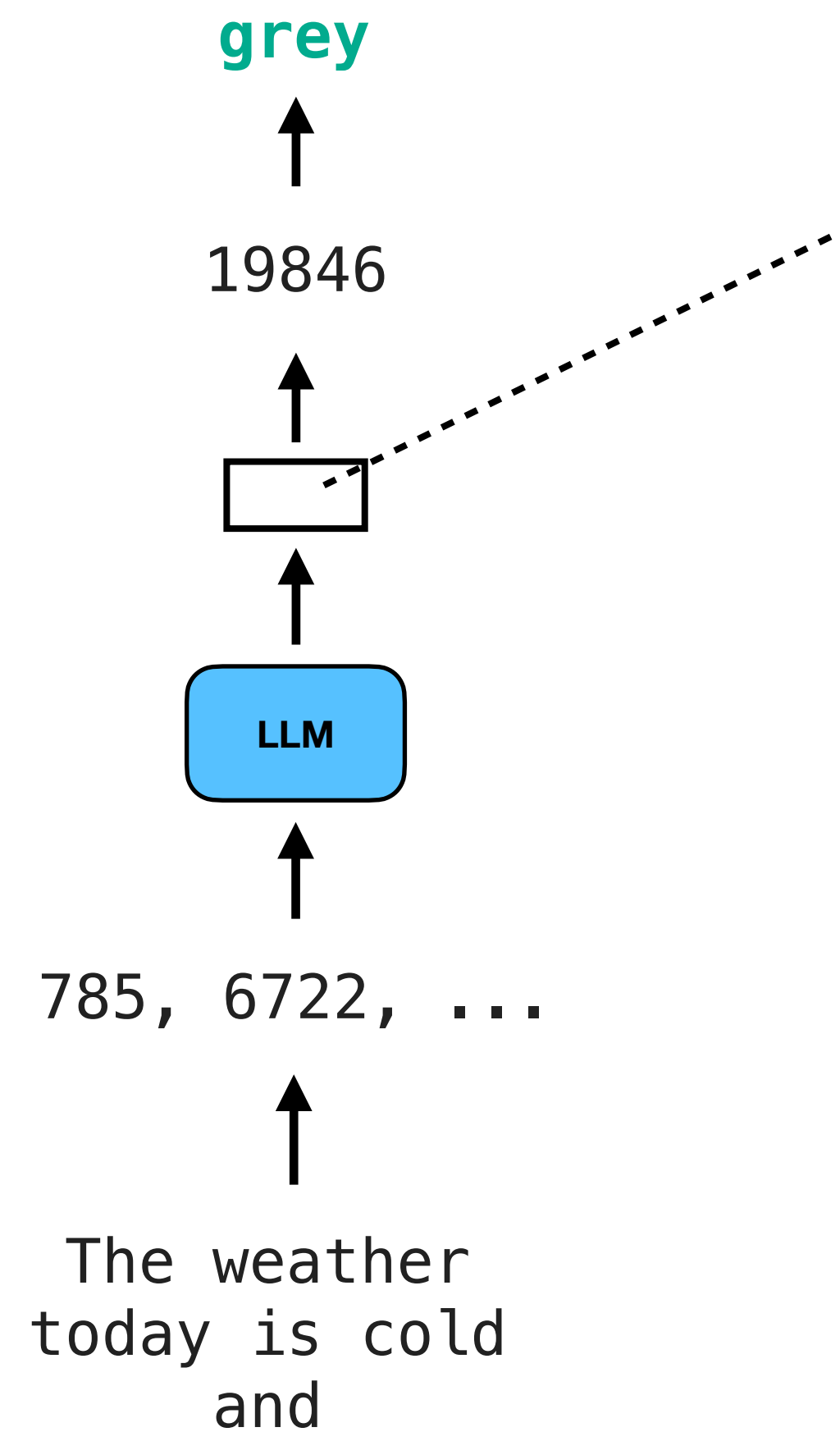
```
    for i in range(5):  
        print(np.random.randint(100))
```

```
[ ]:
```

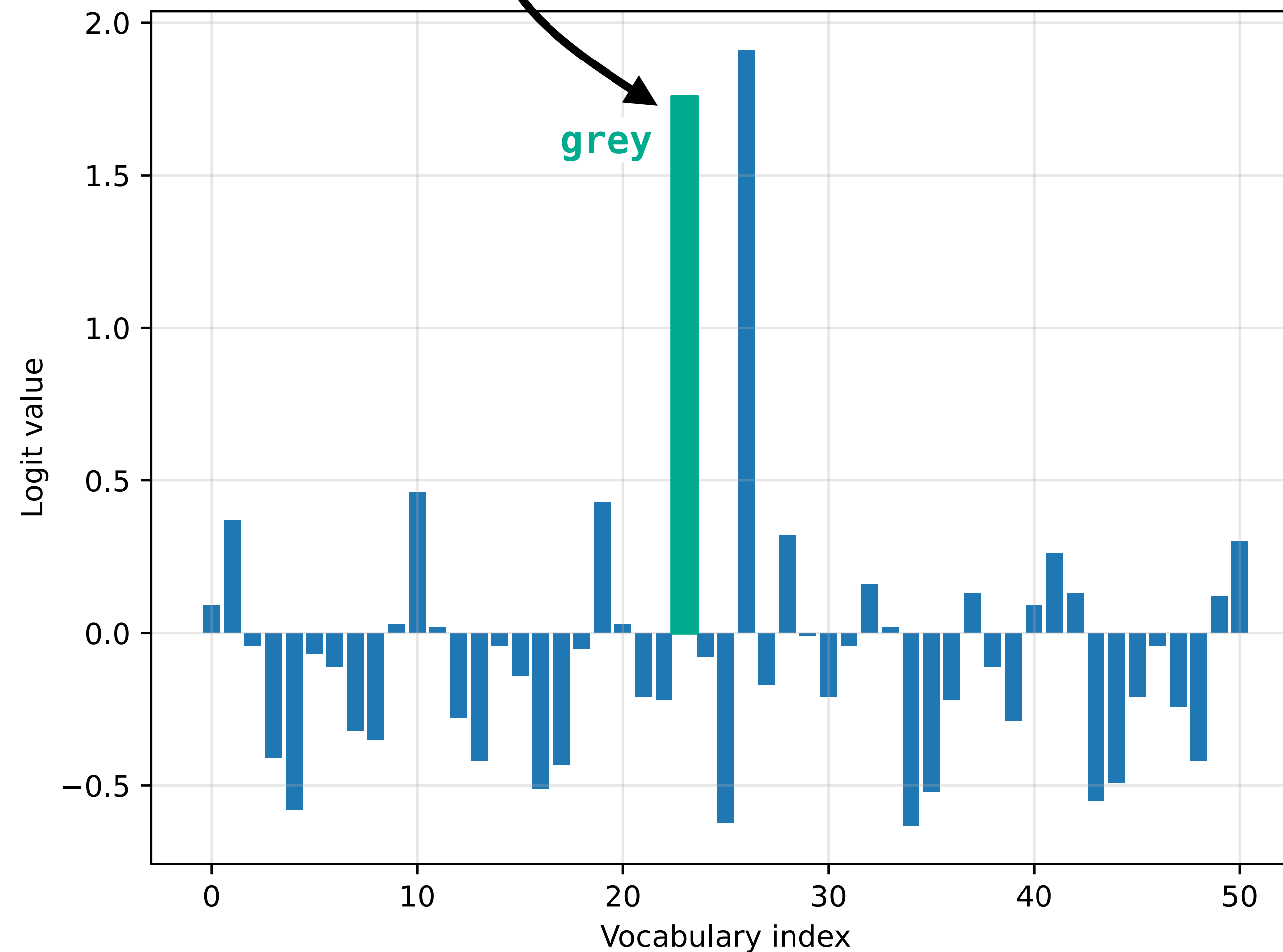
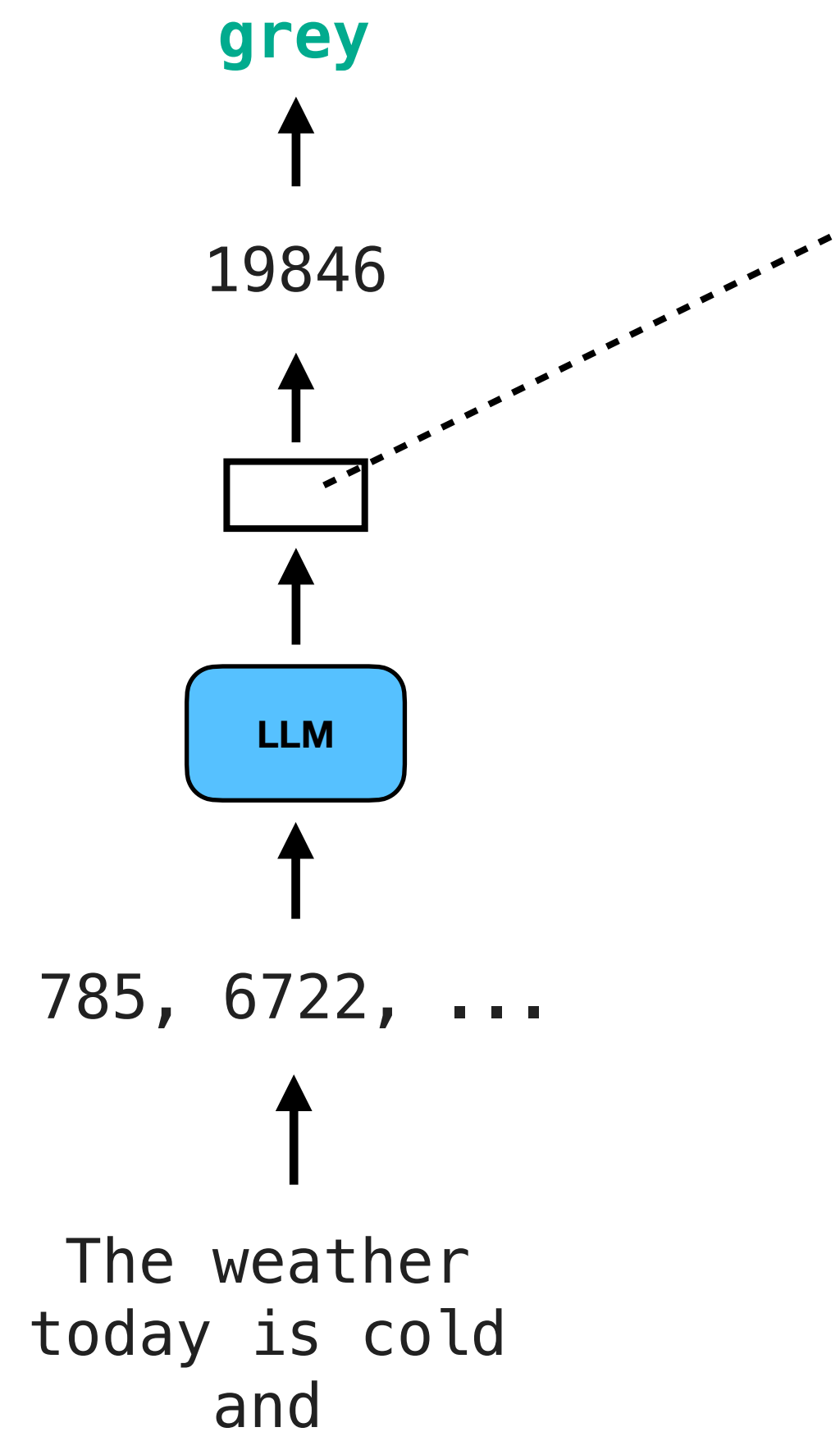
With **random seed 42** we always select this



With **random seed 99** we always select this



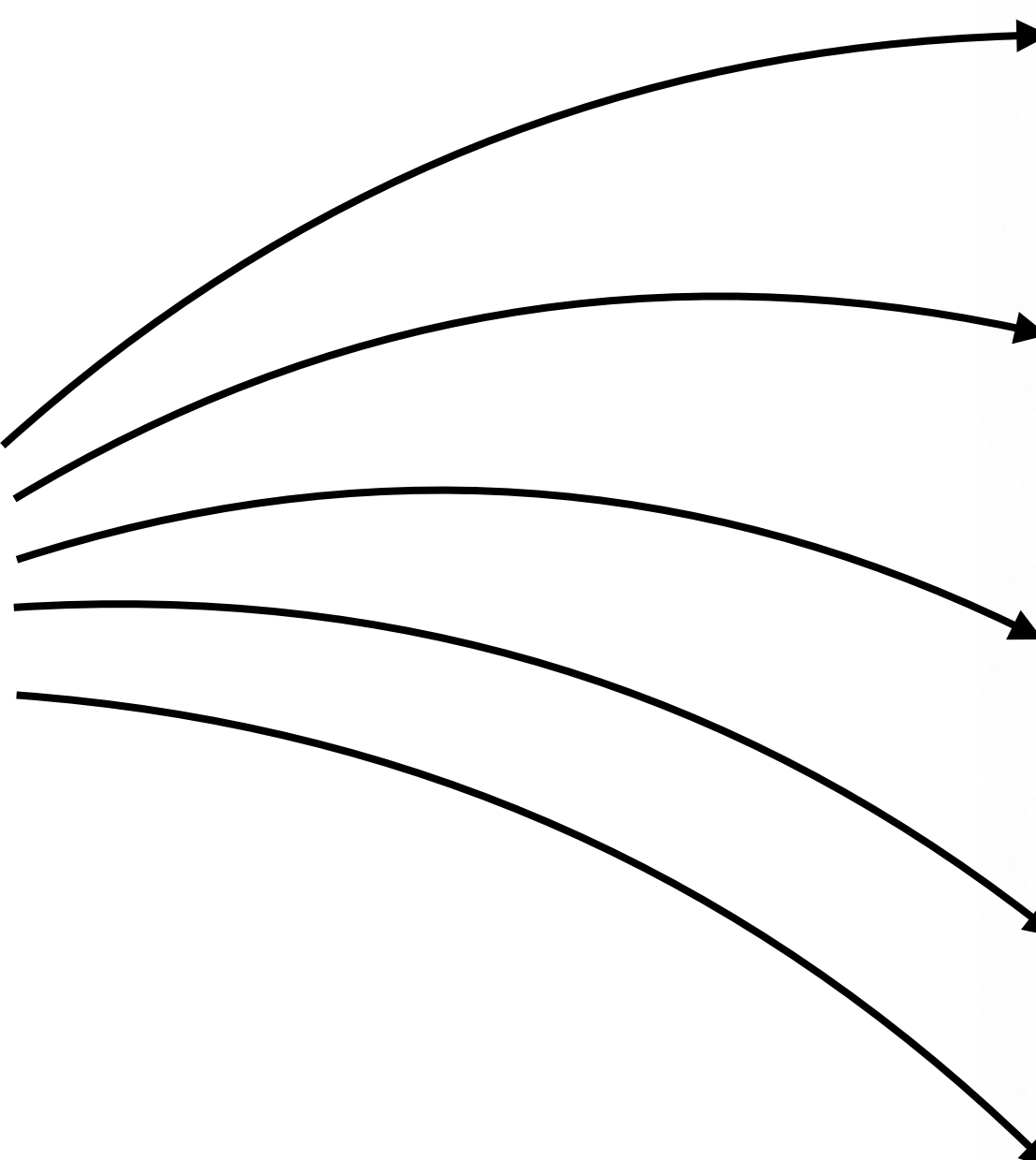
Watermarking sets random seed based on **secret key** (and preceding token context)



Without watermarking might get any of these texts

Plausible texts

The weather today is cold and **overcast** / **grey**. A **light** / **gentle** breeze is **moving** / **blowing** through the trees, and the streets seem **quiet** / **still**. I think I'll stay **home** / **inside** and read a **book** / **novel** with a **cup** / **mug** of tea.

- 
1. The weather today is cold and **overcast**. A **light** breeze is **moving** through the trees, and the streets seem **quiet**. I think I'll stay **home** and read a **book** with a **cup** of tea.
2. The weather today is cold and **grey**. A **gentle** breeze is **blowing** through the trees, and the streets seem **still**. I think I'll stay **inside** and read a **novel** with a **mug** of tea.
3. The weather today is cold and **grey**. A **light** breeze is **blowing** through the trees, and the streets seem **quiet**. I think I'll stay **home** and read a **novel** with a **cup** of tea.
4. The weather today is cold and **overcast**. A **gentle** breeze is **moving** through the trees, and the streets seem **still**. I think I'll stay **inside** and read a **book** with a **mug** of tea.
5. The weather today is cold and **overcast**. A **gentle** breeze is **blowing** through the trees, and the streets seem **quiet**. I think I'll stay **inside** and read a **book** with a **cup** of tea.

...

128. ...

With fixed **random seed 99** we may always get this text

Plausible texts

The weather today is cold and **overcast** / **grey**. A **light** / **gentle** breeze is **moving** / **blowing** through the trees, and the streets seem **quiet** / **still**. I think I'll stay **home** / **inside** and read a **book** / **novel** with a **cup** / **mug** of tea.

1. The weather today is cold and **overcast**. A **light** breeze is **moving** through the trees, and the streets seem **quiet**. I think I'll stay **home** and read a **book** with a **cup** of tea.

2. The weather today is cold and **grey**. A **gentle** breeze is **blowing** through the trees, and the streets seem **still**. I think I'll stay **inside** and read a **novel** with a **mug** of tea.

3. The weather today is cold and **grey**. A **light** breeze is **blowing** through the trees, and the streets seem **quiet**. I think I'll stay **home** and read a **novel** with a **cup** of tea.

4. The weather today is cold and **overcast**. A **gentle** breeze is **moving** through the trees, and the streets seem **still**. I think I'll stay **inside** and read a **book** with a **mug** of tea.

5. The weather today is cold and **overcast**. A **gentle** breeze is **blowing** through the trees, and the streets seem **quiet**. I think I'll stay **inside** and read a **book** with a **cup** of tea.

...

128. ...

With watermarking we might get this text

Secret key = 8F3A91C7...

Plausible texts

The weather today is cold and **overcast** / **grey**. A **light** / **gentle** breeze is **moving** / **blowing** through the trees, and the streets seem **quiet** / **still**. I think I'll stay **home** / **inside** and read a **book** / **novel** with a **cup** / **mug** of tea.

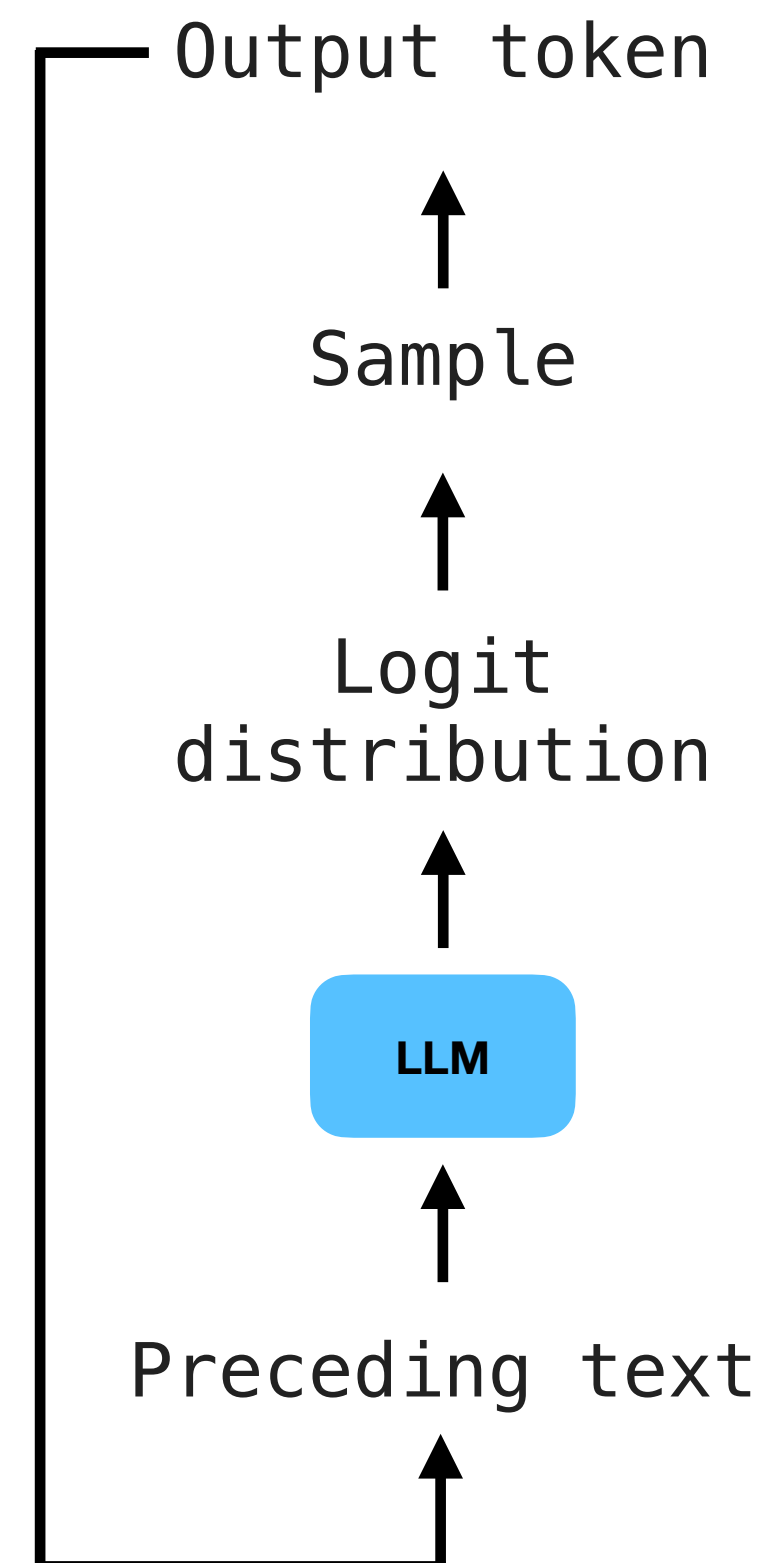
-
1. The weather today is cold and **overcast**. A **light** breeze is **moving** through the trees, and the streets seem **quiet**. I think I'll stay **home** and read a **book** with a **cup** of tea.
 2. The weather today is cold and **grey**. A **gentle** breeze is **blowing** through the trees, and the streets seem **still**. I think I'll stay **inside** and read a **novel** with a **mug** of tea.
 3. The weather today is cold and **grey**. A **light** breeze is **blowing** through the trees, and the streets seem **quiet**. I think I'll stay **home** and read a **novel** with a **cup** of tea.
 4. The weather today is cold and **overcast**. A **gentle** breeze is **moving** through the trees, and the streets seem **still**. I think I'll stay **inside** and read a **book** with a **mug** of tea.
 5. The weather today is cold and **overcast**. A **gentle** breeze is **blowing** through the trees, and the streets seem **quiet**. I think I'll stay **inside** and read a **book** with a **cup** of tea.

...

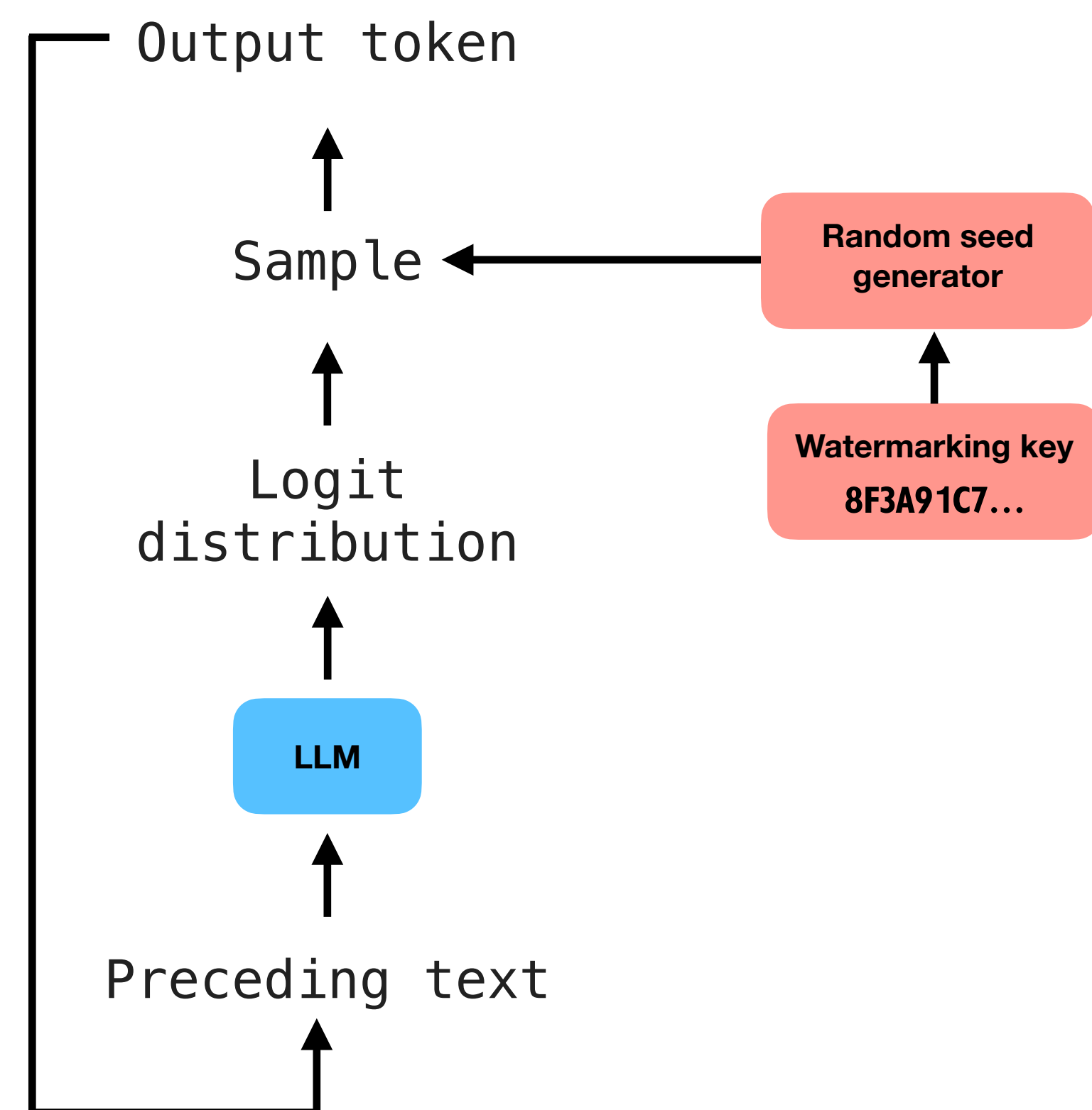
128. ...

Summary

No watermarking



With watermarking

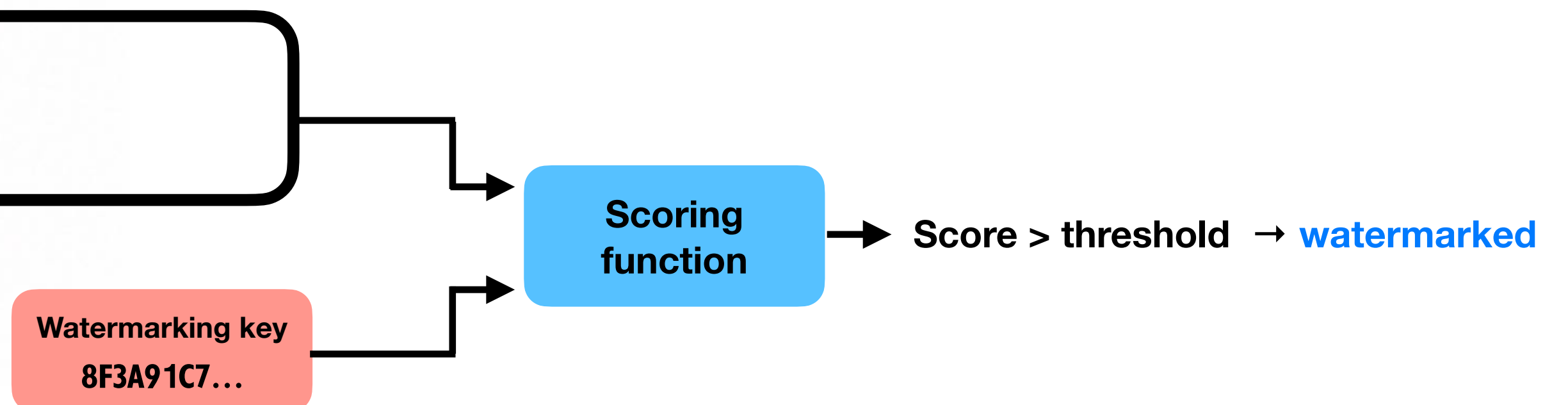


Detect watermark: Only possible if access to key

1. The weather today is cold and **overcast**. A **light** breeze is **moving** through the trees, and the streets seem **quiet**. I think I'll stay **home** and read a **book** with a **cup** of tea.
2. The weather today is cold and **grey**. A **gentle** breeze is **blowing** through the trees, and the streets seem **still**. I think I'll stay **inside** and read a **novel** with a **mug** of tea.
3. The weather today is cold and **grey**. A **light** breeze is **blowing** through the trees, and the streets seem **quiet**. I think I'll stay **home** and read a **novel** with a **cup** of tea.
4. The weather today is cold and **overcast**. A **gentle** breeze is **moving** through the trees, and the streets seem **still**. I think I'll stay **inside** and read a **book** with a **mug** of tea.
5. The weather today is cold and **overcast**. A **gentle** breeze is **blowing** through the trees, and the streets seem **quiet**. I think I'll stay **inside** and read a **book** with a **cup** of tea.

...

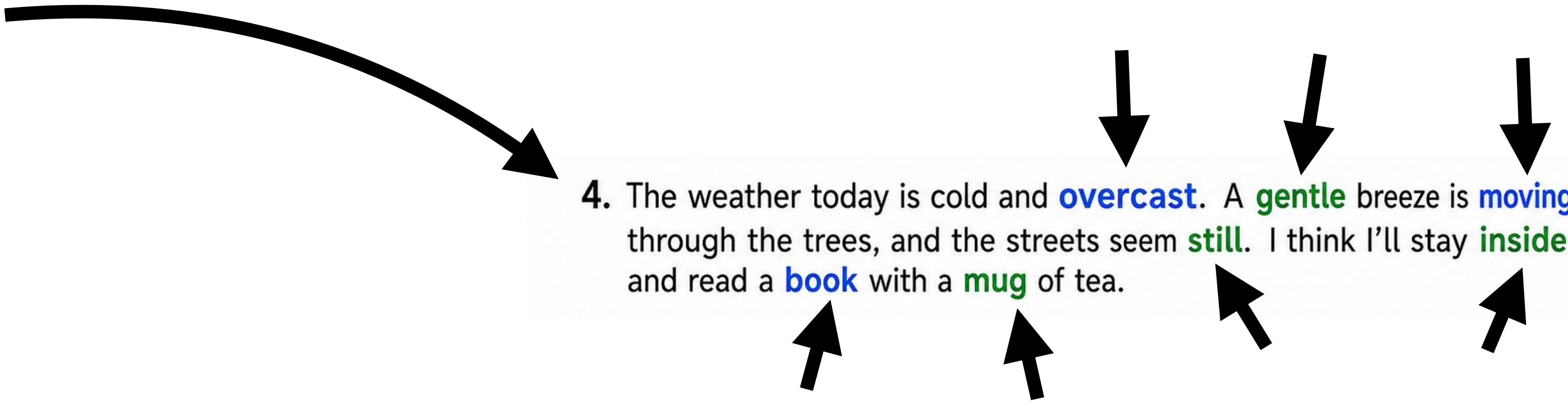
128. ...



Remove watermark by editing words

Plausible texts

The weather today is cold and
overcast / **grey**. A **light** / **gentle**
breeze is **moving** / **blowing** through
the trees, and the streets seem
quiet / **still**. I think I'll stay
home / **inside** and read a
book / **novel** with a **cup** / **mug** of tea.



A diagram illustrating the process of removing a watermark by editing words. A large curved arrow points from the 'Plausible texts' section to a numbered list item. From this list item, four straight arrows point to specific words: 'overcast', 'gentle', 'moving', and 'inside'.

4. The weather today is cold and **overcast**. A **gentle** breeze is **moving**
through the trees, and the streets seem **still**. I think I'll stay **inside**
and read a **book** with a **mug** of tea.

Remove watermark by editing words

Plausible texts

The weather today is cold and
overcast / **grey**. A **light** / **gentle**
breeze is **moving** / **blowing** through
the trees, and the streets seem
quiet / **still**. I think I'll stay
home / **inside** and read a
book / **novel** with a **cup** / **mug** of tea.



4. The weather today is cold and **overcast**. A **gentle** breeze is **moving**
through the trees, and the streets seem **still**. I think I'll stay **inside**
and read a **book** with a **mug** of tea.

now (orange) points to **overcast**
cloudy (orange) points to **overcast**
blowing (green) points to **moving**
novel (green) points to **book**
cup (blue) points to **mug**
guess (orange) points to **inside**

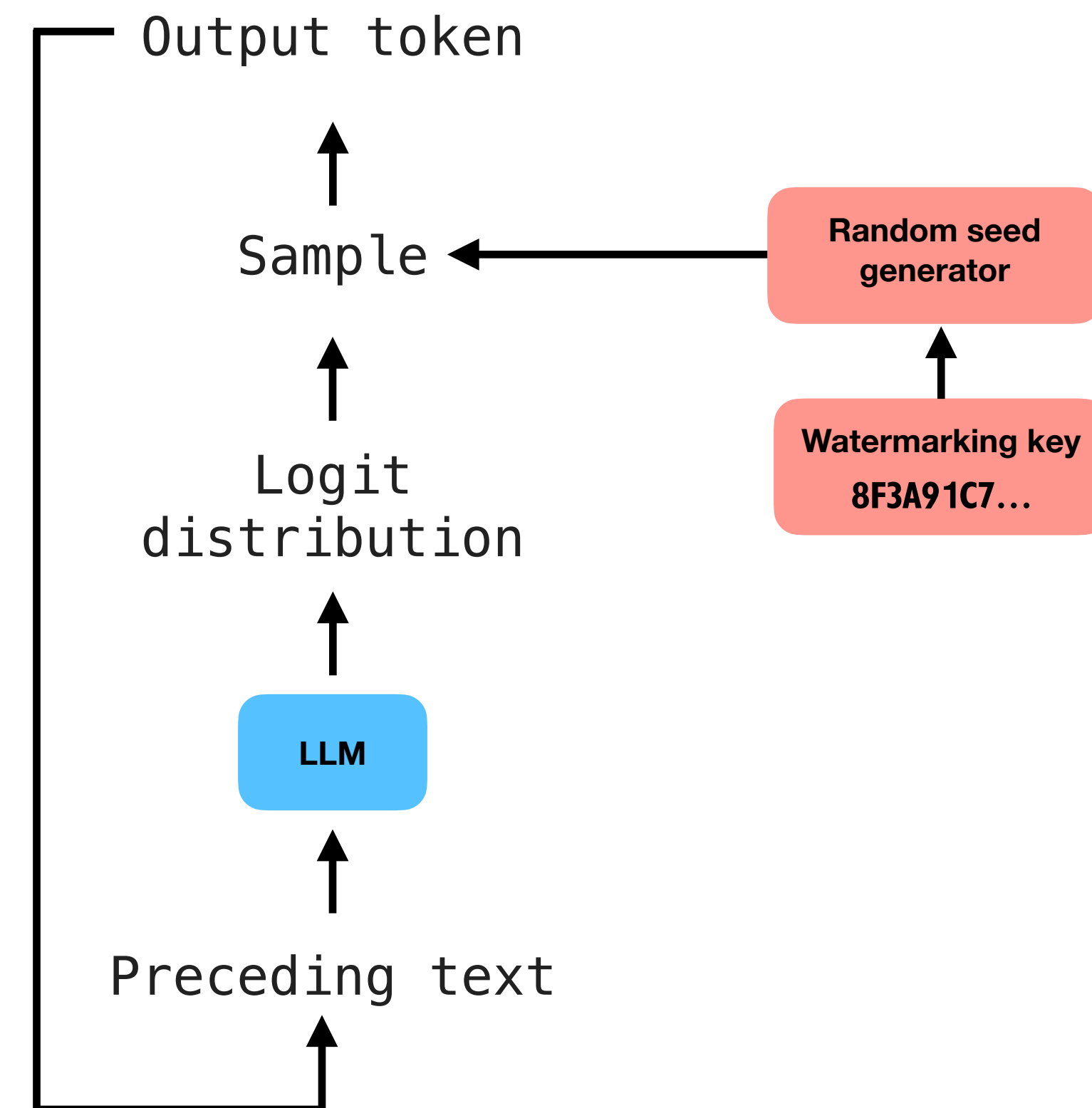
Bonus: How does the **scoring** function work?

SynthID-Text method from <https://www.nature.com/articles/s41586-024-08025-4>

Bonus: How does the **scoring** function work?

SynthID-Text method from <https://www.nature.com/articles/s41586-024-08025-4>

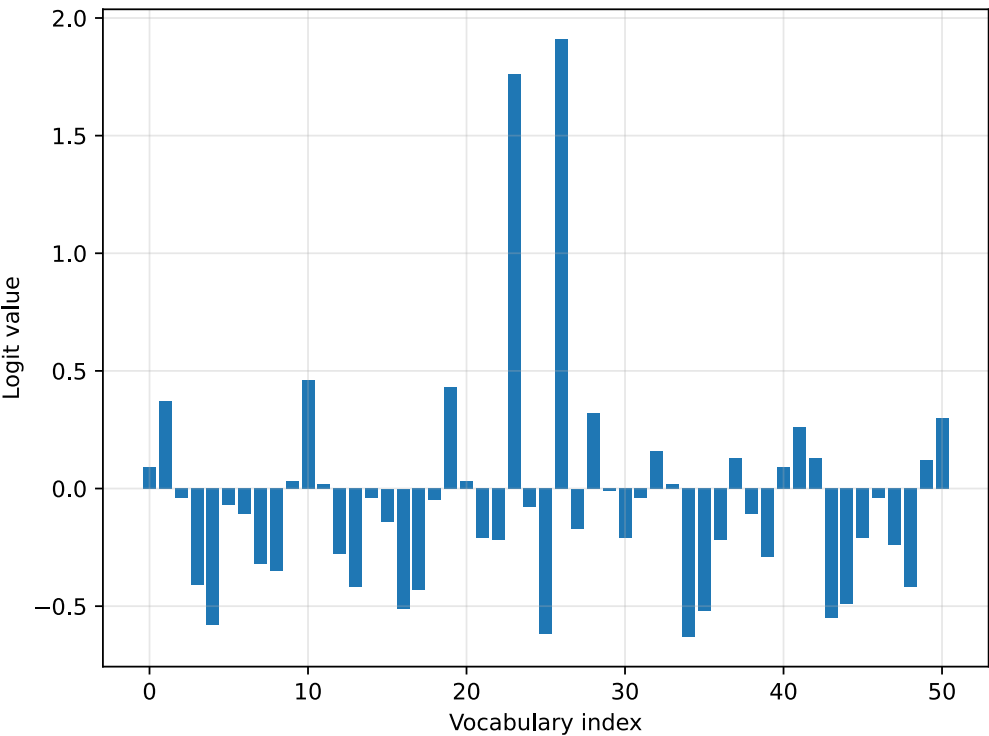
During generation there is a bit more nuance how the next token is sampled:



Bonus: How does the **scoring** function work?

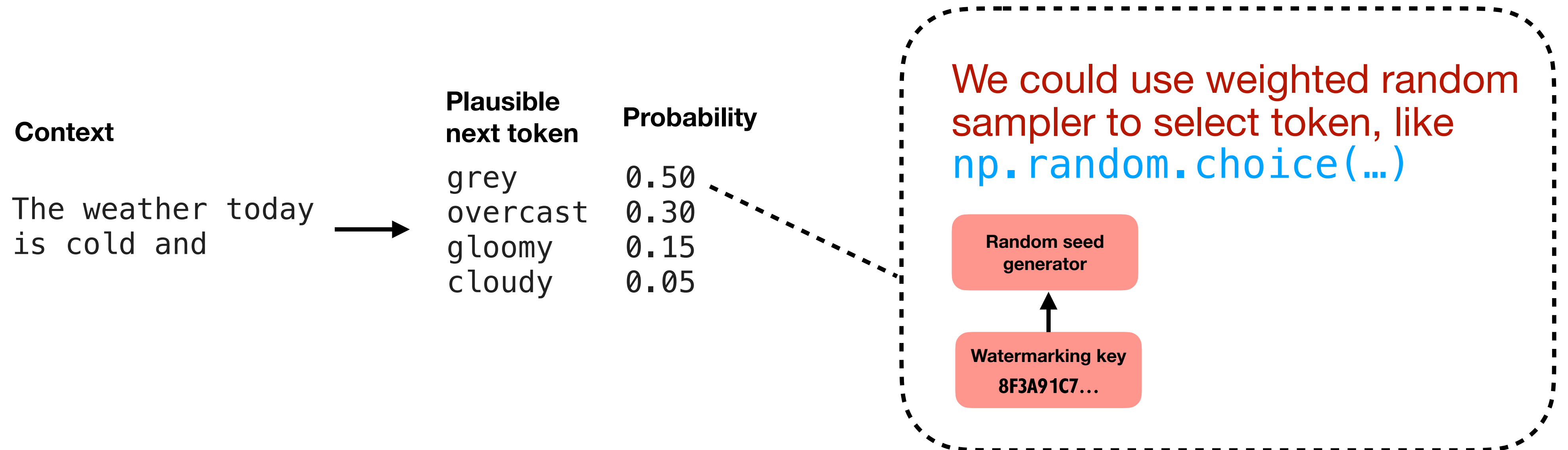
SynthID-Text method from <https://www.nature.com/articles/s41586-024-08025-4>

Context		Plausible next token	Probability
The weather today is cold and	→	grey	0.50
		overcast	0.30
		gloomy	0.15
		cloudy	0.05



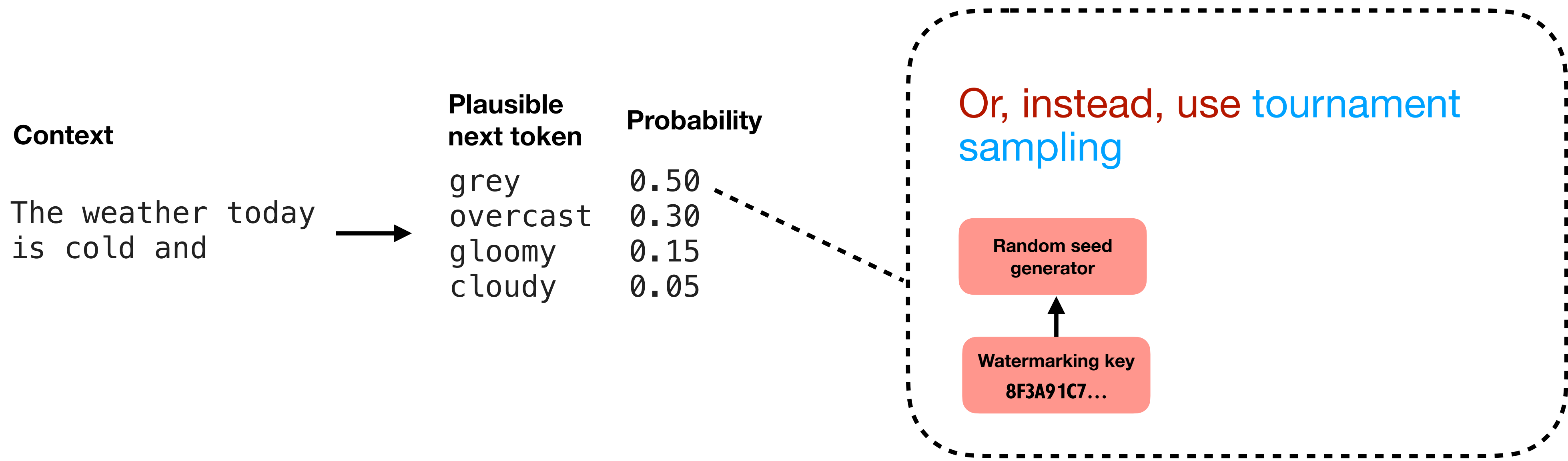
Bonus: How does the **scoring** function work?

SynthID-Text method from <https://www.nature.com/articles/s41586-024-08025-4>



Bonus: How does the **scoring** function work?

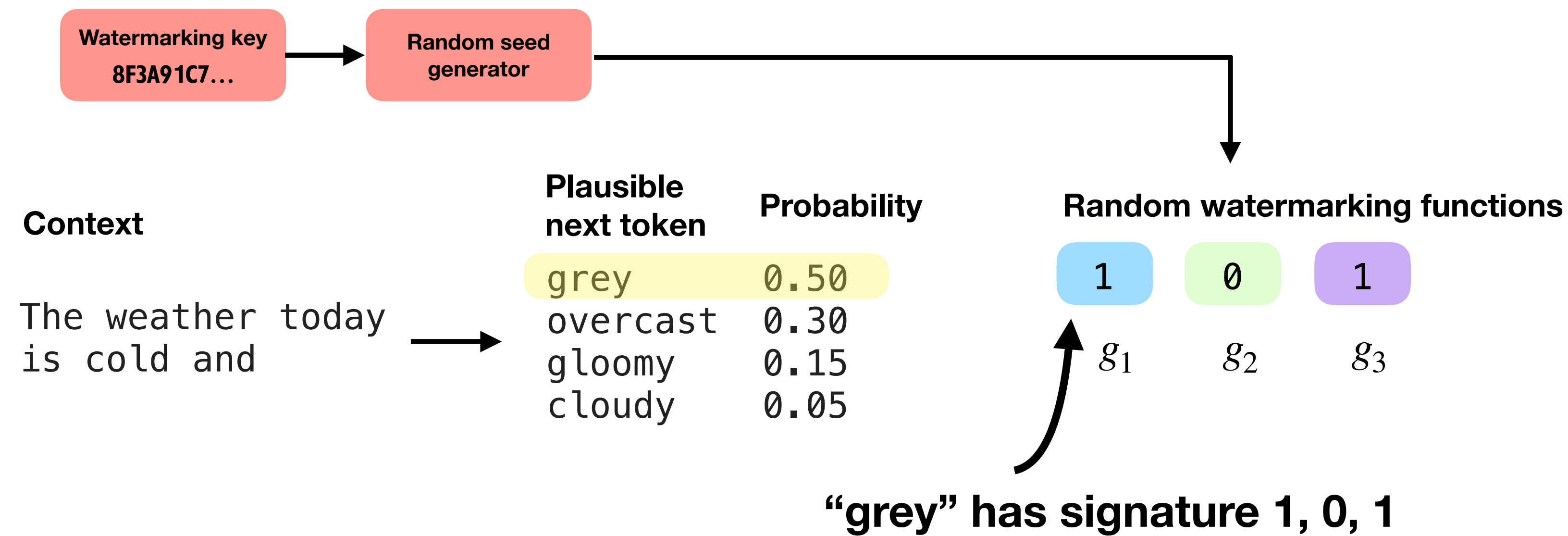
SynthID-Text method from <https://www.nature.com/articles/s41586-024-08025-4>



Bonus: How does the **scoring** function work?

SynthID-Text method from <https://www.nature.com/articles/s41586-024-08025-4>

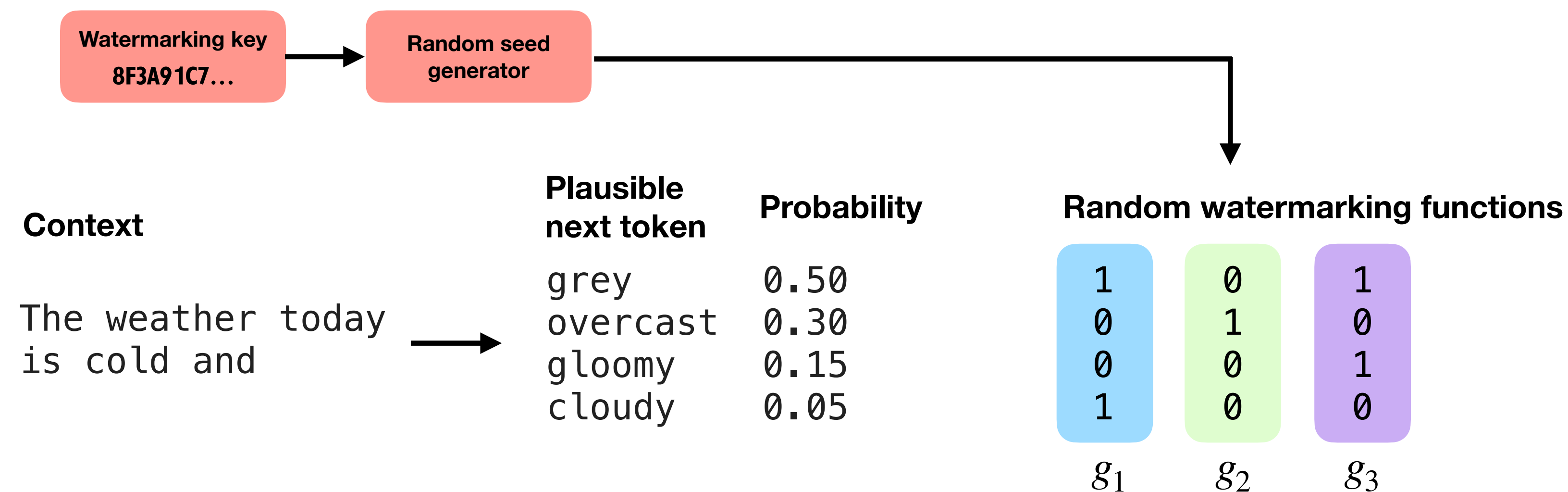
Generate bit assignment with random watermark functions



Bonus: How does the **scoring** function work?

SynthID-Text method from <https://www.nature.com/articles/s41586-024-08025-4>

Generate bit assignment with random watermark functions



Bonus: How does the **scoring** function work?

SynthID-Text method from <https://www.nature.com/articles/s41586-024-08025-4>

Tournament sampling:

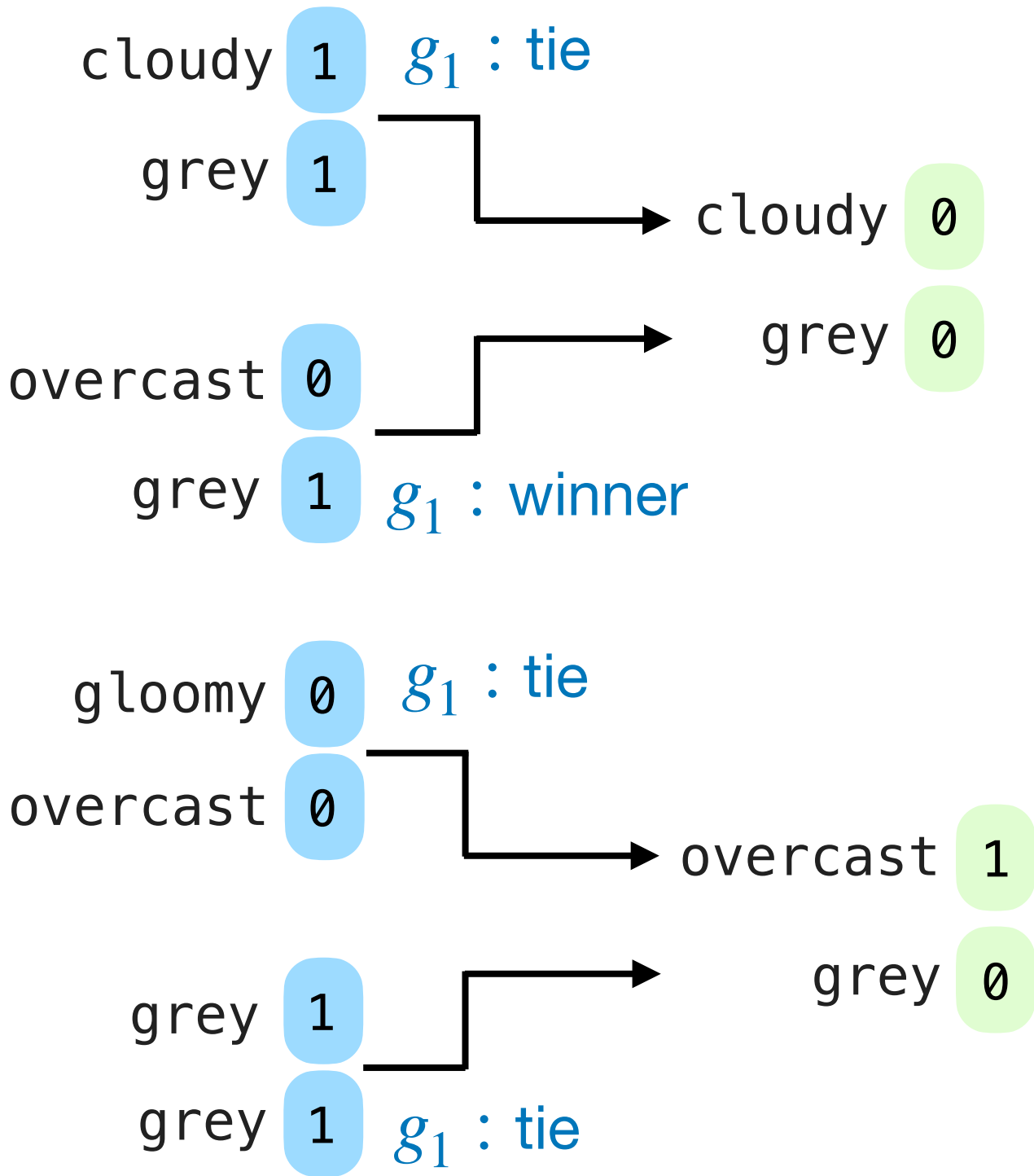
cloudy	1
grey	1
overcast	0
grey	1
gloomy	0
overcast	0
grey	1
grey	1

grey	0.50	1	0	1
overcast	0.30	0	1	0
gloomy	0.15	0	0	1
cloudy	0.05	1	0	0
		g_1	g_2	g_3

Bonus: How does the **scoring** function work?

SynthID-Text method from <https://www.nature.com/articles/s41586-024-08025-4>

Tournament sampling:

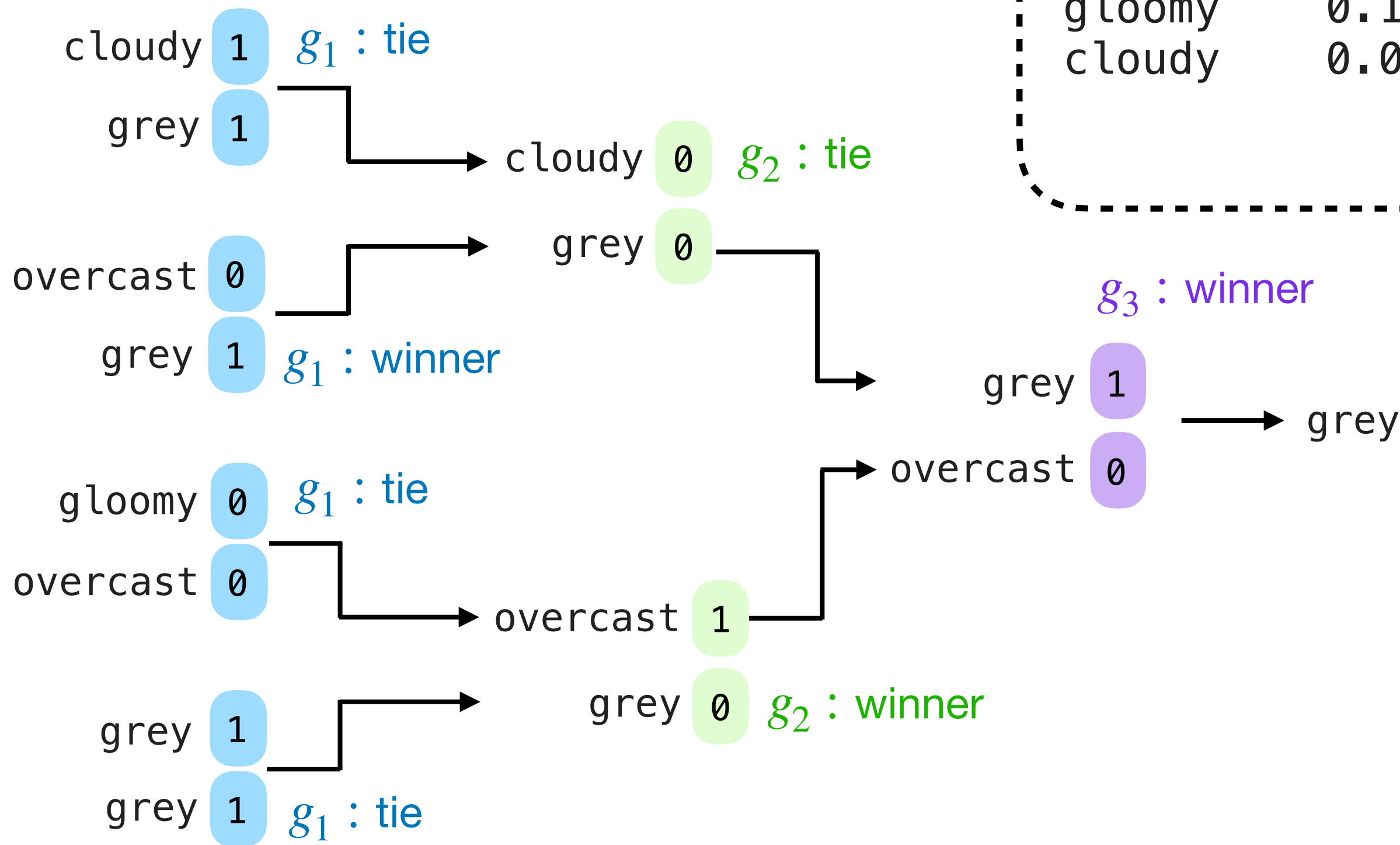


grey	0.50	1	0	1
overcast	0.30	0	1	0
gloomy	0.15	0	0	1
cloudy	0.05	1	0	0
		g_1	g_2	g_3

Bonus: How does the **scoring** function work?

SynthID-Text method from <https://www.nature.com/articles/s41586-024-08025-4>

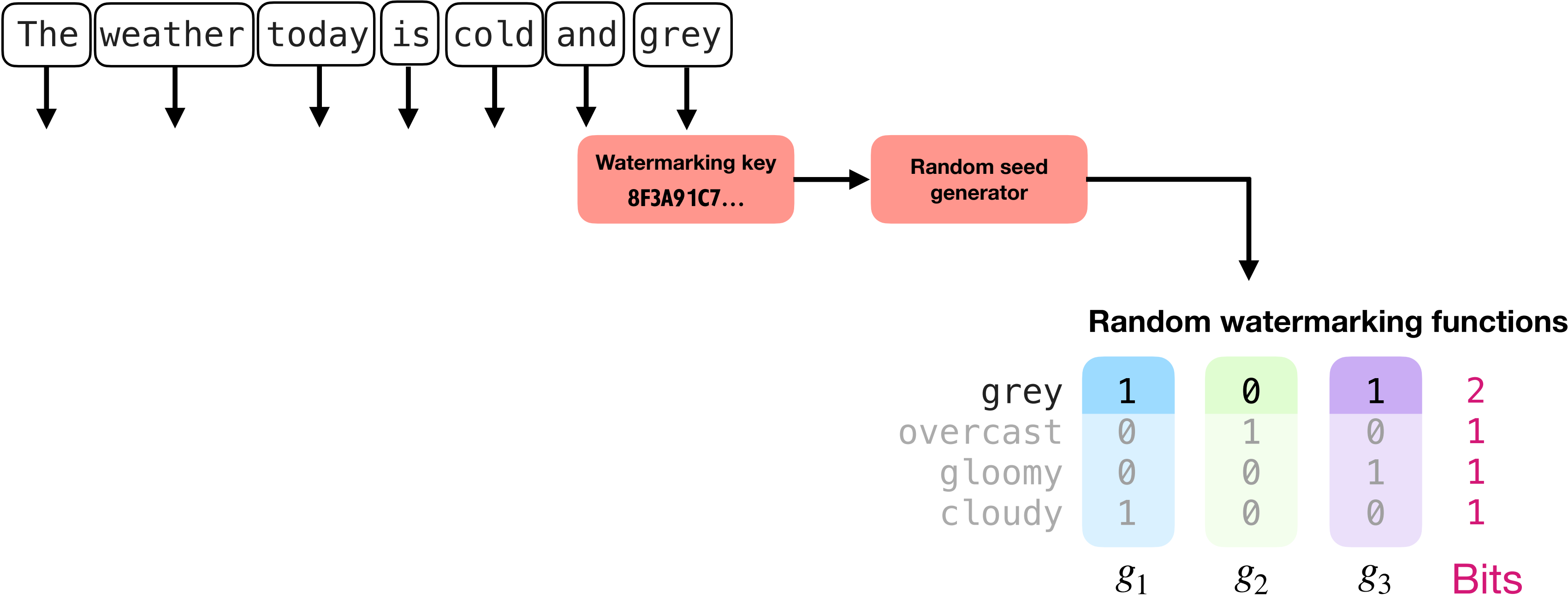
Tournament sampling:



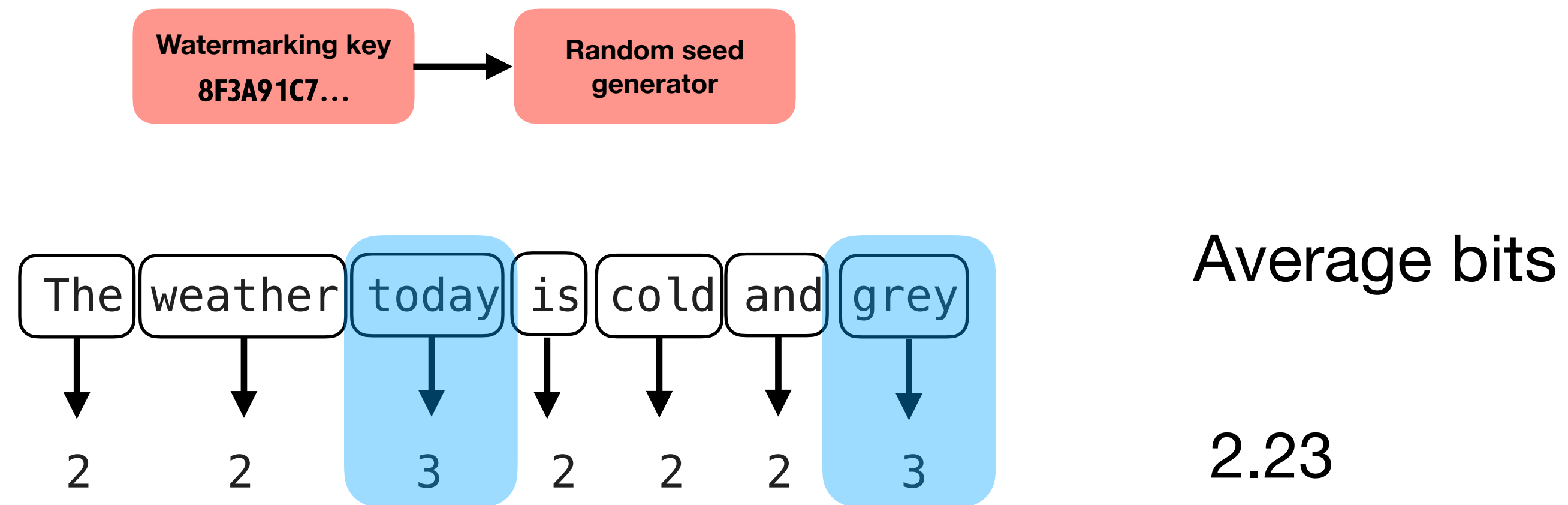
grey	0.50	1	0	1
overcast	0.30	0	1	0
gloomy	0.15	0	0	1
cloudy	0.05	1	0	0
		g_1	g_2	g_3

**Tournament sampling is
mainly to make detection
easier (cheaper)**

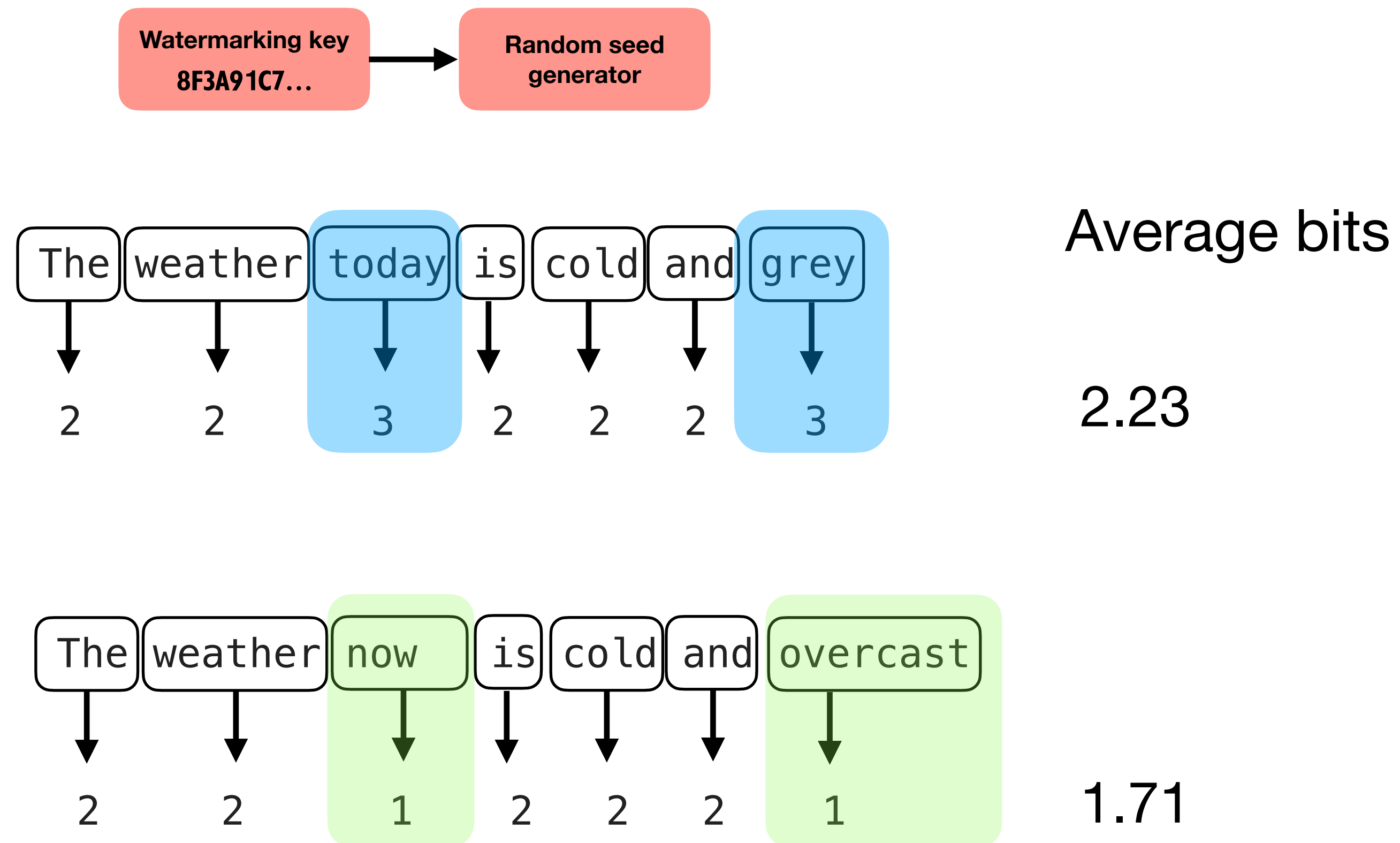
After sampling (during scoring)



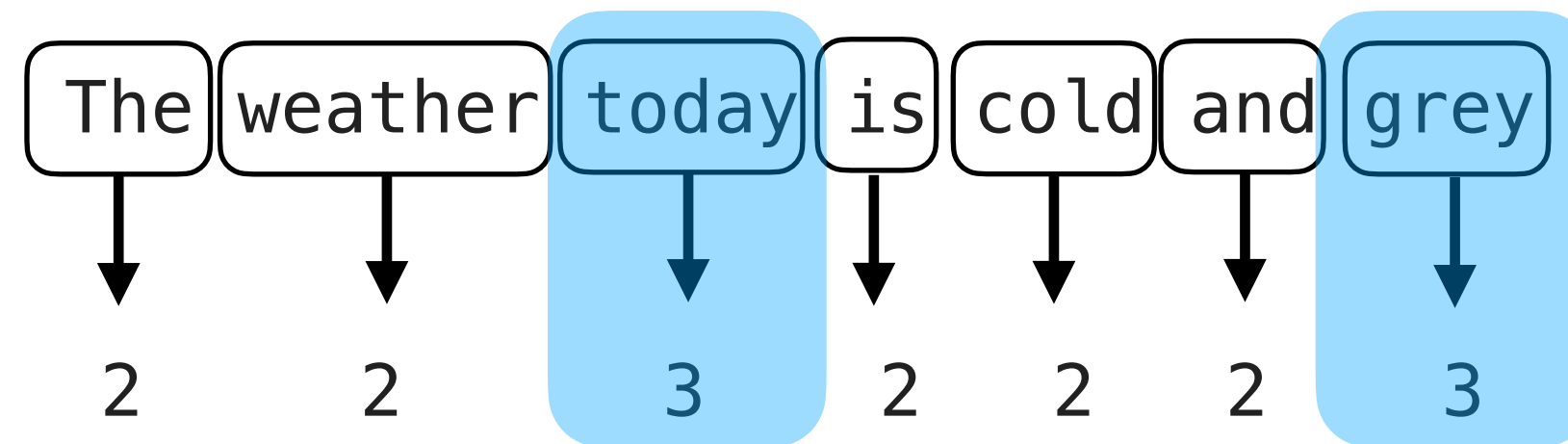
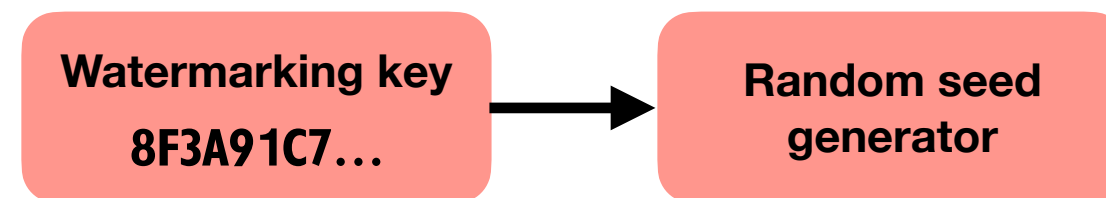
After sampling (during scoring)



After sampling (during scoring)

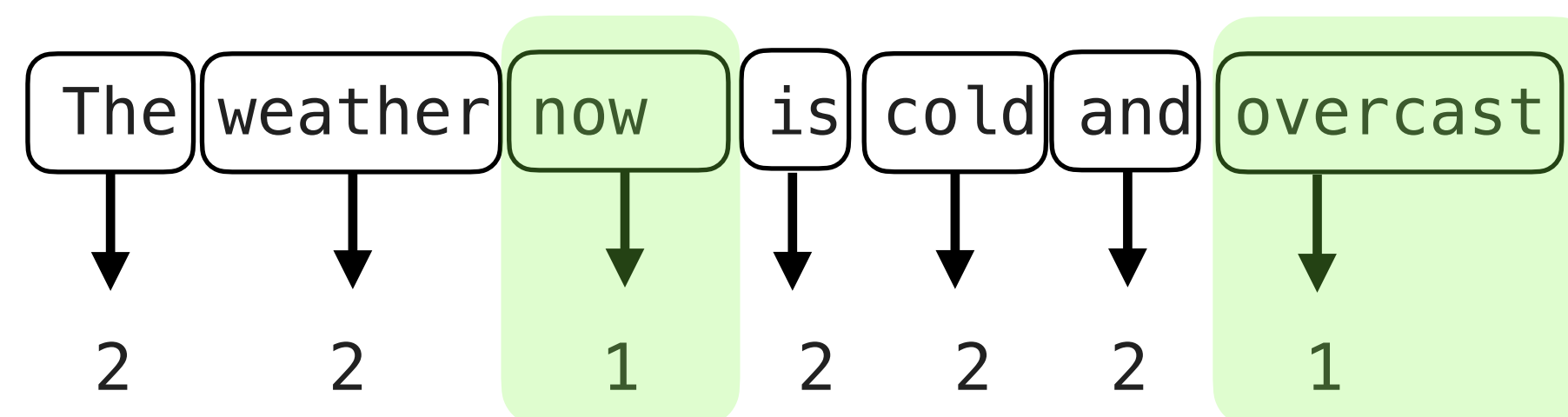


After sampling (during scoring)



Average bits

2.23 → if score > 2 then watermarked

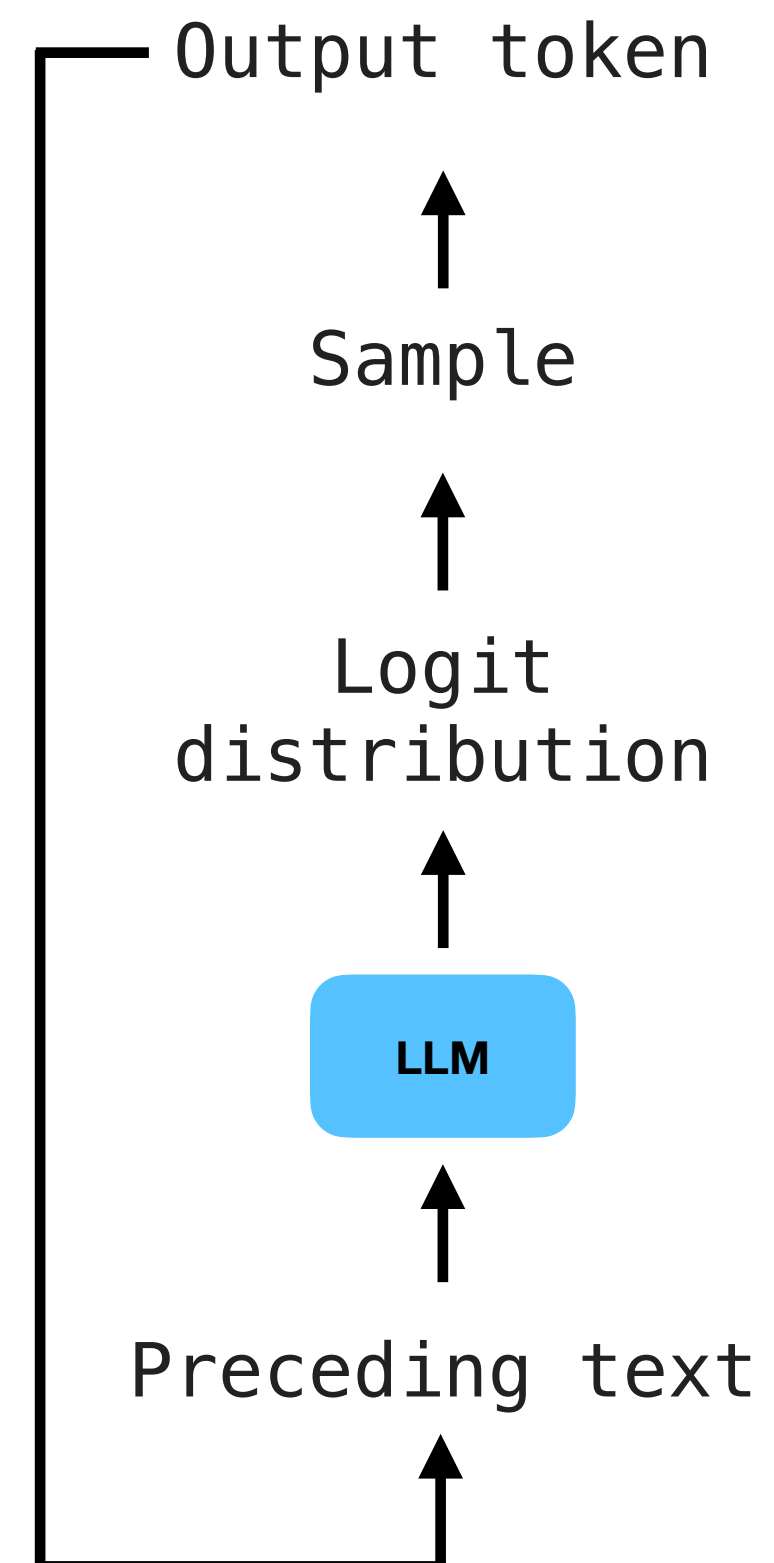


1.71 → if score > 2 then watermarked

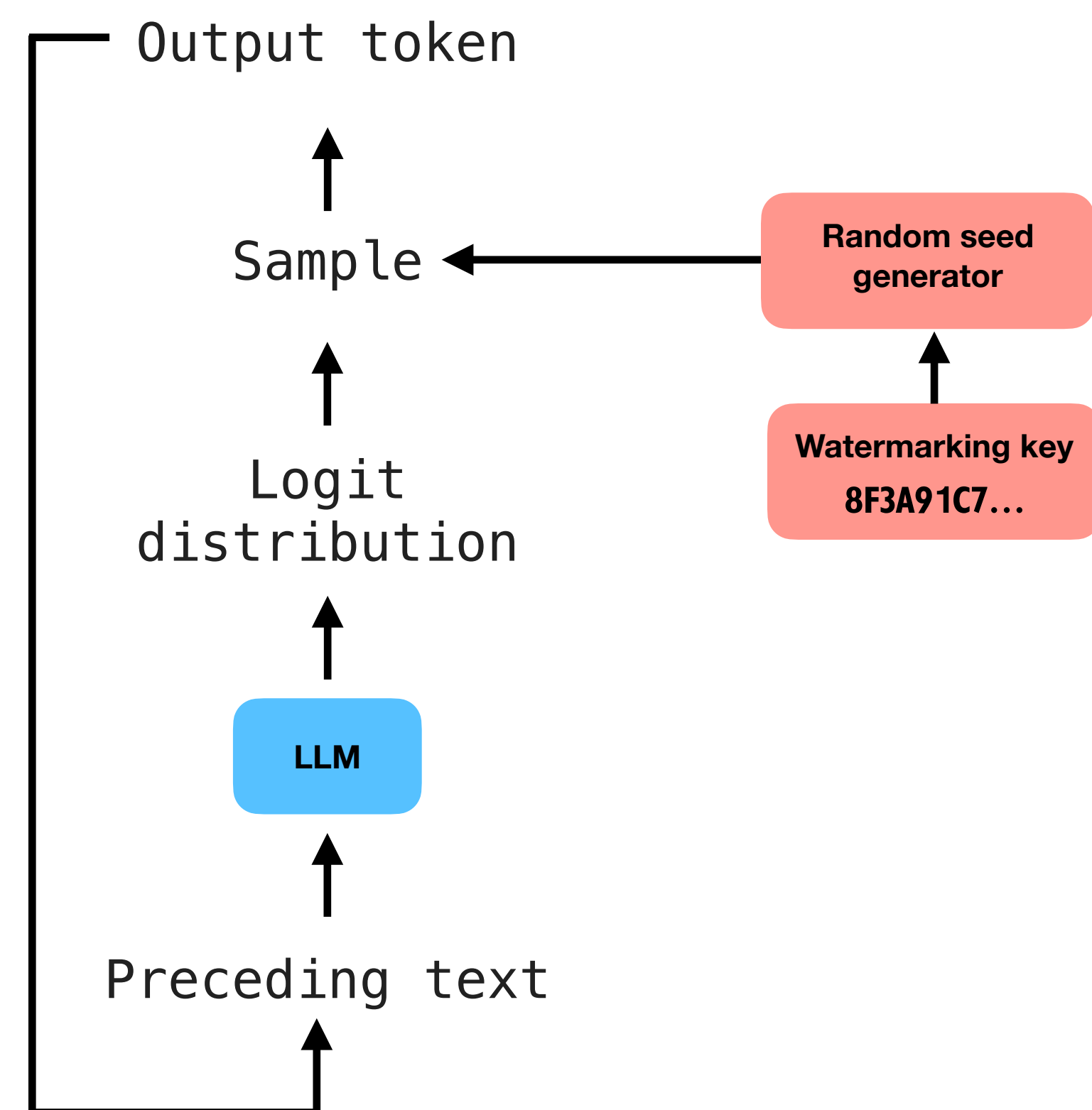
**Again, tournament sampling
is mainly to make detection
easier (cheaper)**

Summary

No watermarking



With watermarking



**Detecting watermarks requires
secret key and watermarking
functions $(g_1 \dots g_n)$**

Remove watermark by editing words

Plausible texts

The weather today is cold and **overcast** / **grey**. A **light** / **gentle** breeze is **moving** / **blowing** through the trees, and the streets seem **quiet** / **still**. I think I'll stay **home** / **inside** and read a **book** / **novel** with a **cup** / **mug** of tea.

ideally, edit all watermark positions:

4. The weather today is cold and **overcast**. A **gentle** breeze is **moving** through the trees, and the streets seem **still**. I think I'll stay **inside** and read a **book** with a **mug** of tea.

Remove watermark by editing words

Plausible texts

The weather today is cold and **overcast** / **grey**. A **light** / **gentle** breeze is **moving** / **blowing** through the trees, and the streets seem **quiet** / **still**. I think I'll stay **home** / **inside** and read a **book** / **novel** with a **cup** / **mug** of tea.

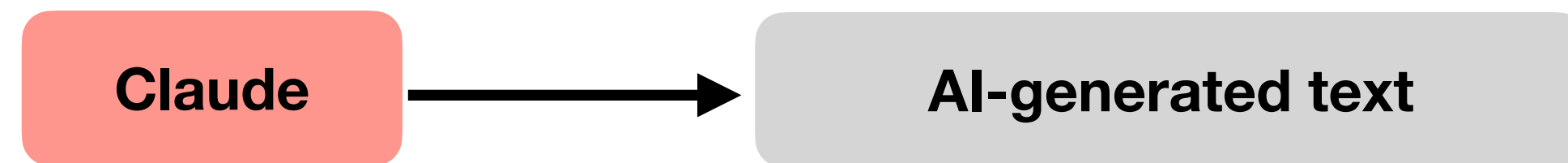
In practice, edit some random positions:

4. The weather today is cold and **overcast**. A **gentle** breeze is **moving** through the trees, and the streets seem **still**. I think I'll stay **inside** and read a **book** with a **mug** of tea.

Diagram illustrating word editing in the text:

- now** (orange) points to **overcast** (blue).
- cloudy** (orange) points to **overcast** (blue).
- blowing** (green) points to **moving** (blue).
- novel** (green) points to **book** (blue).
- cup** (blue) points to **mug** (green).
- guess** (orange) points to **inside** (green).

The future of AI-generated text is **worse** AI-generated text?



The future of AI-generated text is worse AI-generated text?

