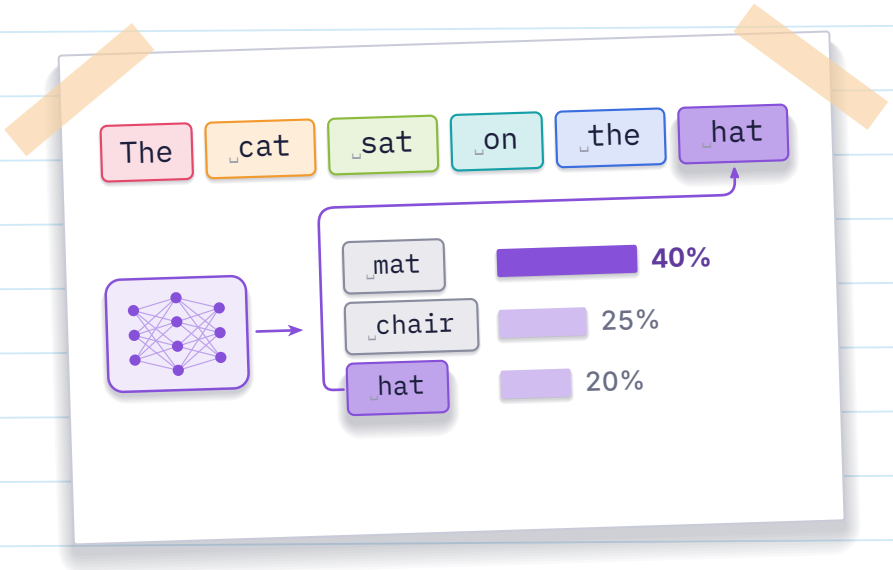
Prof. Aviv Zohar

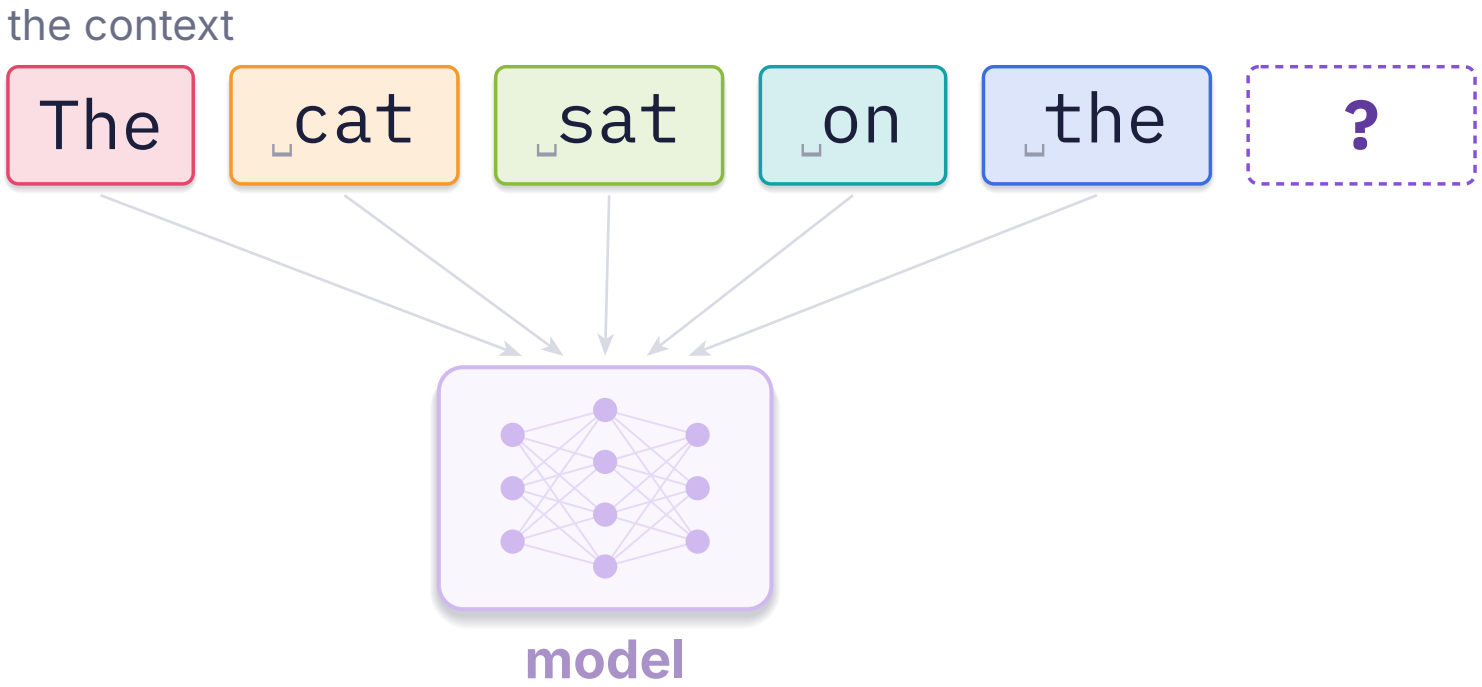
Semester A, 2026–7

LLMs: From Predictions to a Chatting Assistant

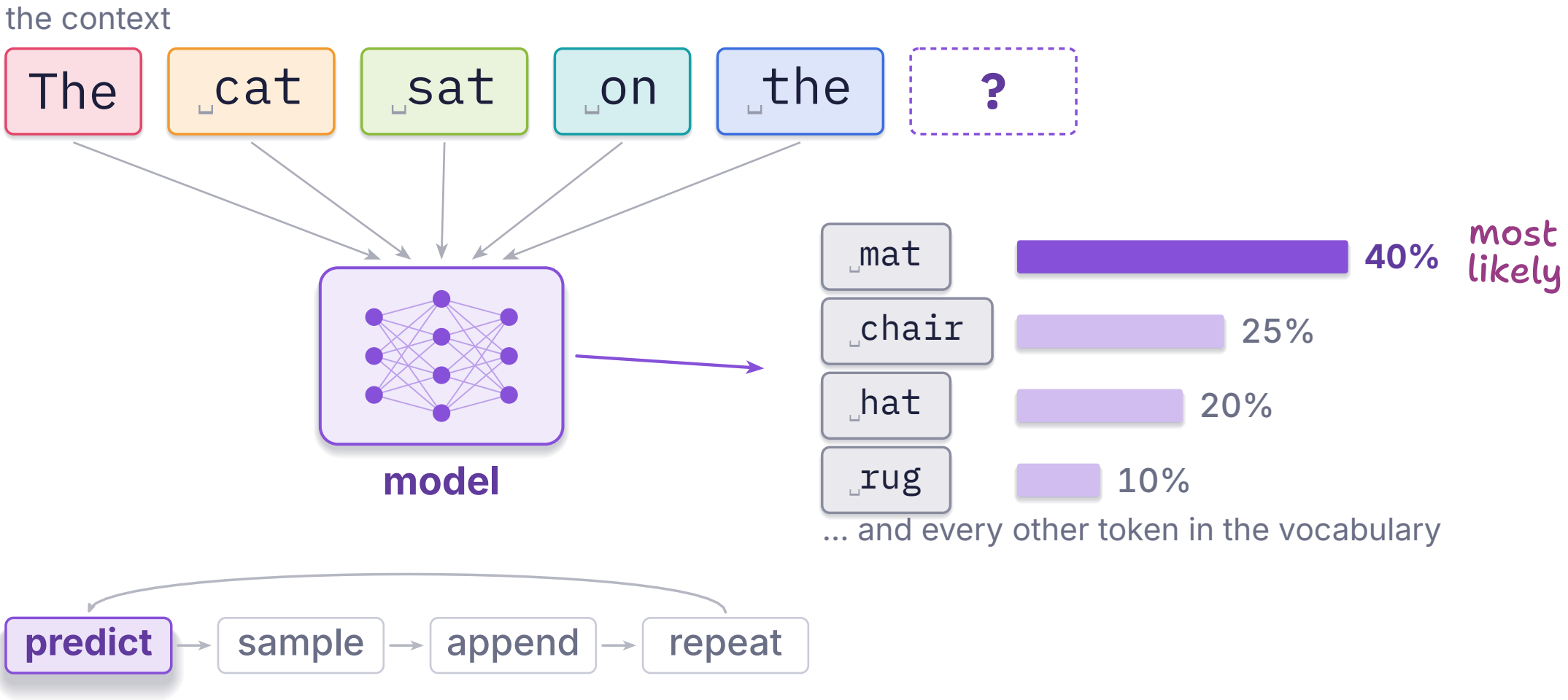
Sampling, chat and the context window



Text is generated one token at a time

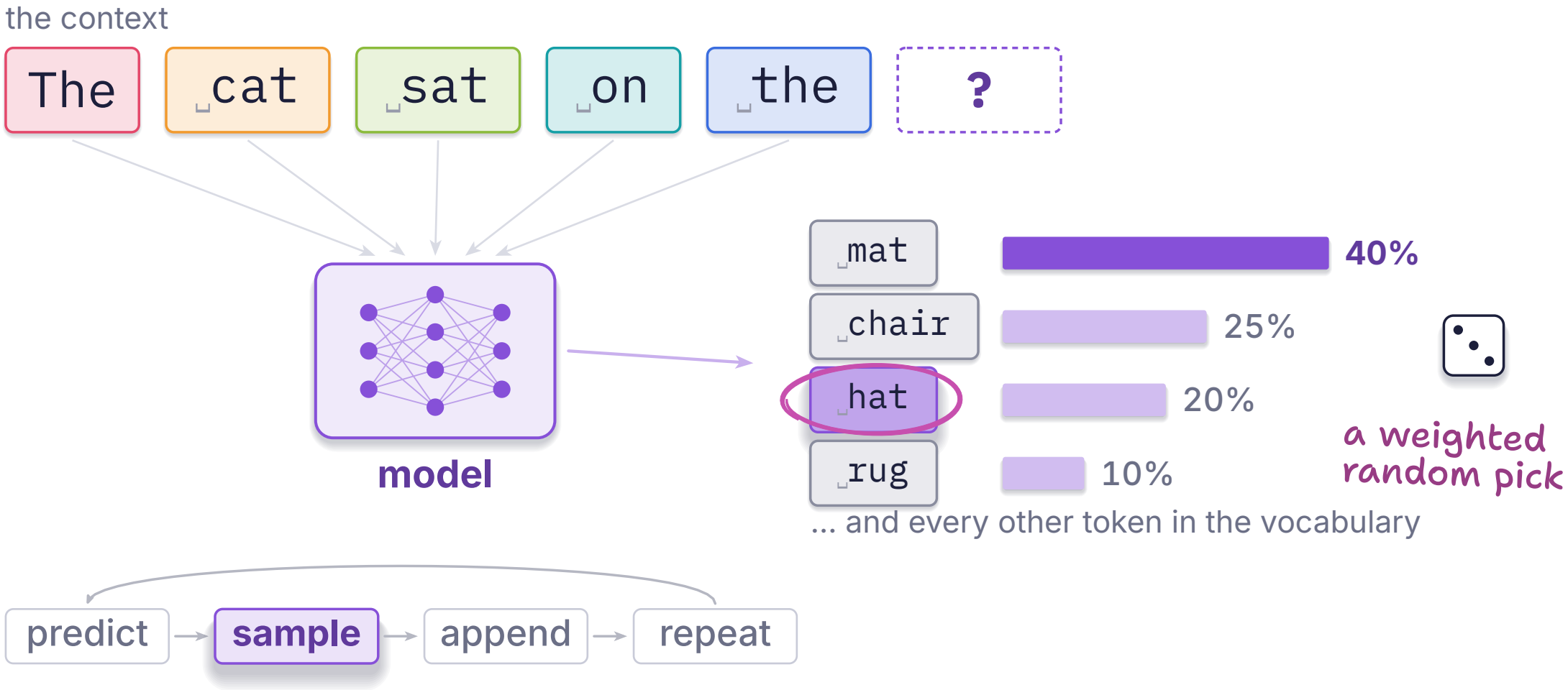


Text is generated one token at a time



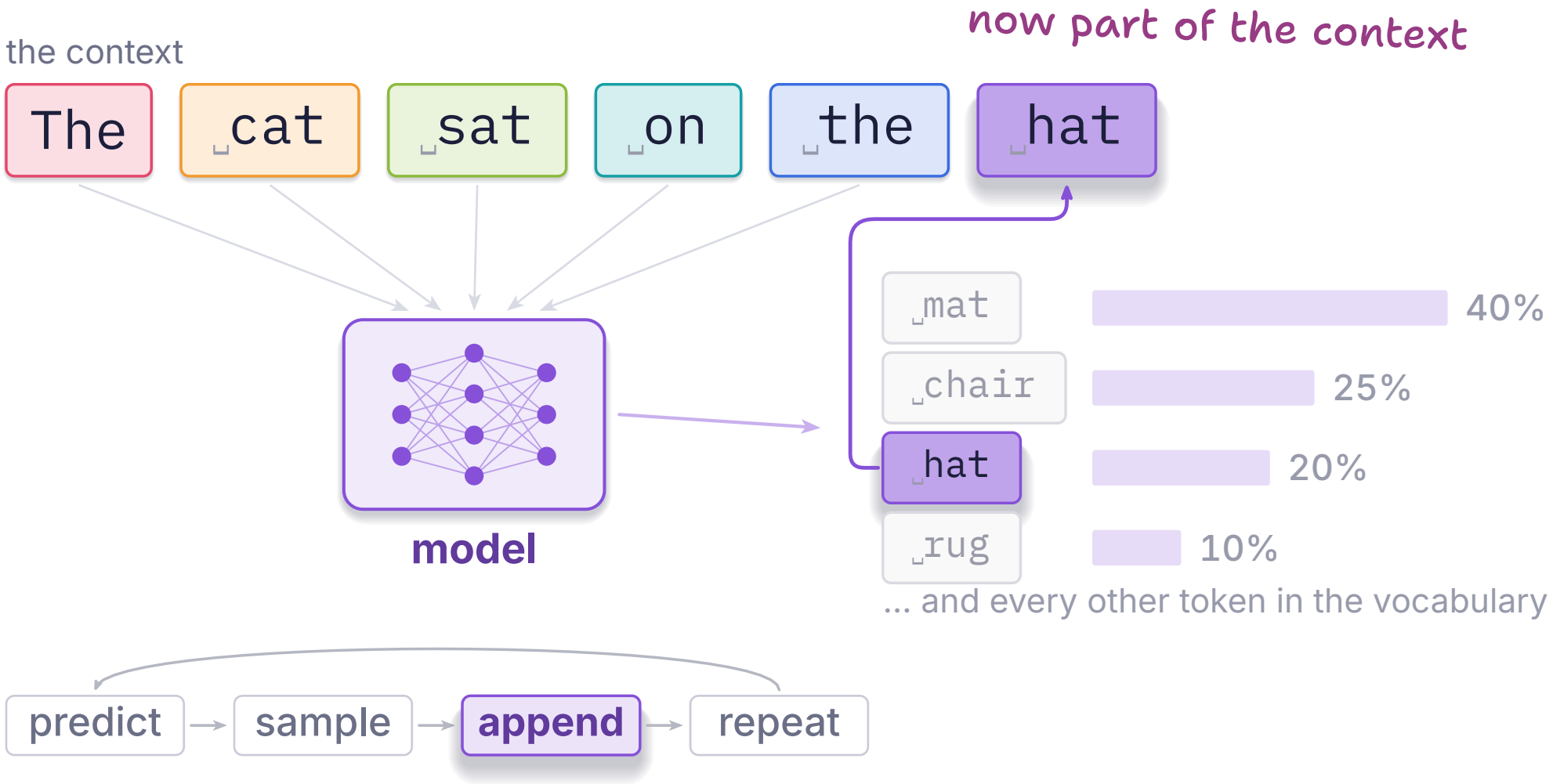
probabilities are illustrative

Text is generated one token at a time



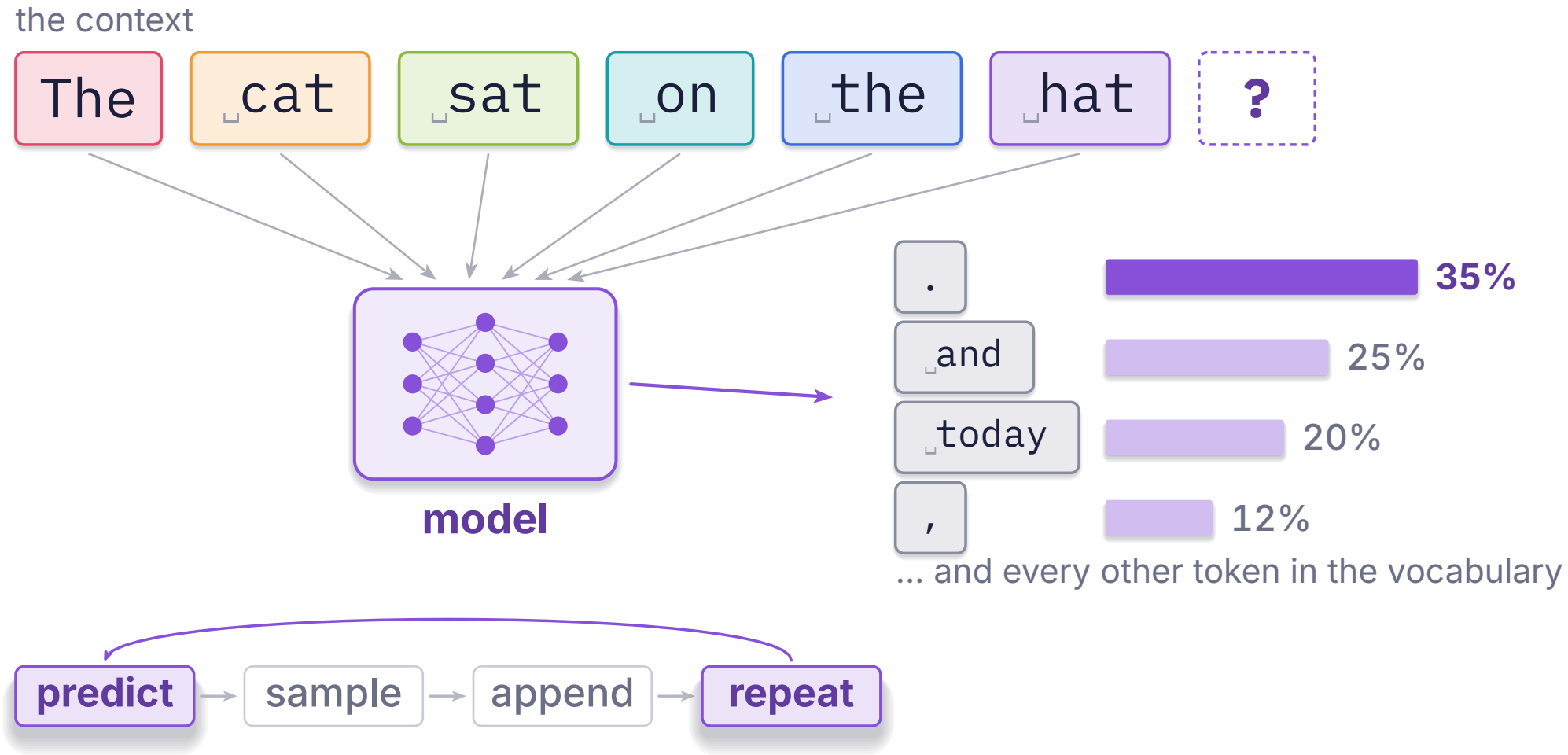
probabilities are illustrative

Text is generated one token at a time



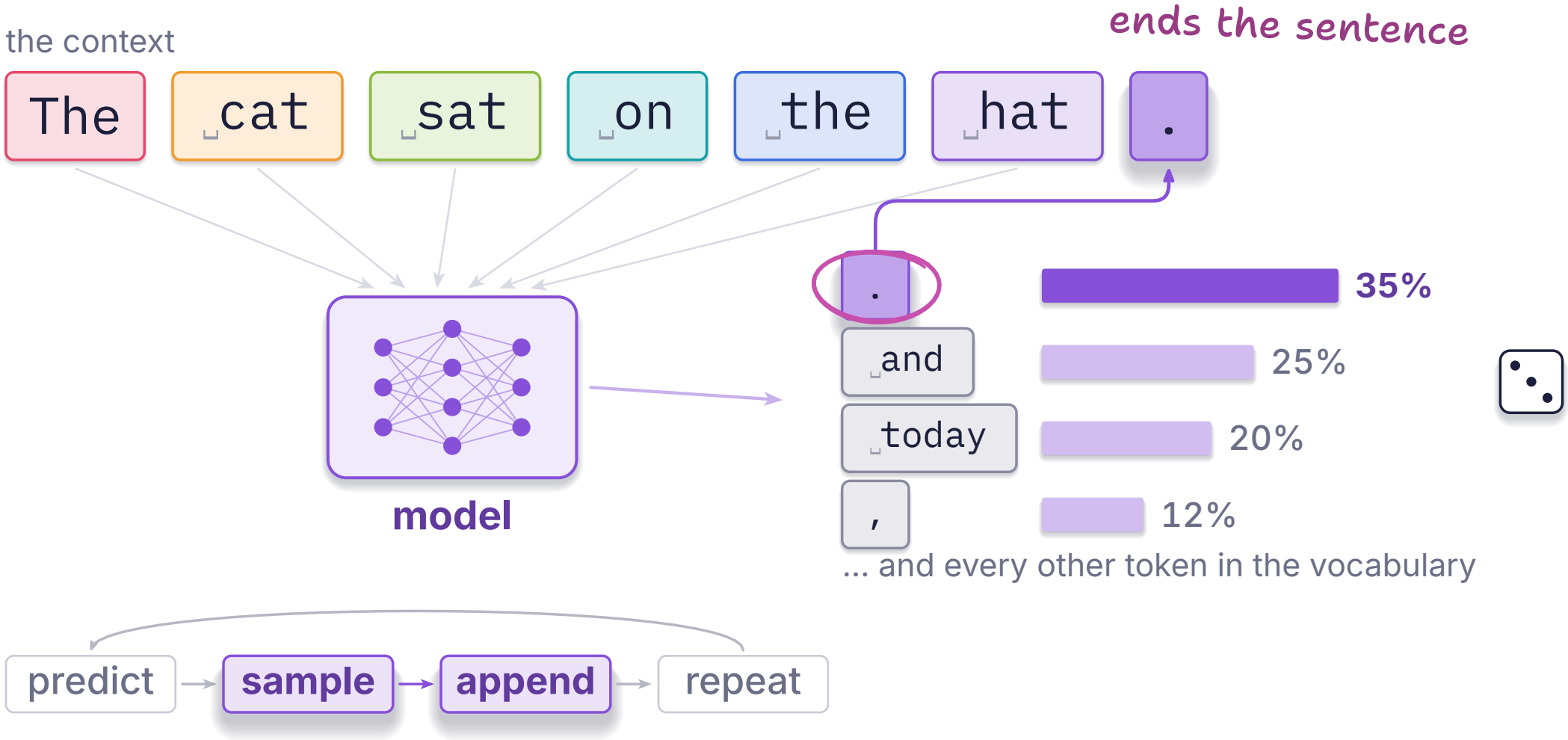
probabilities are illustrative

Text is generated one token at a time



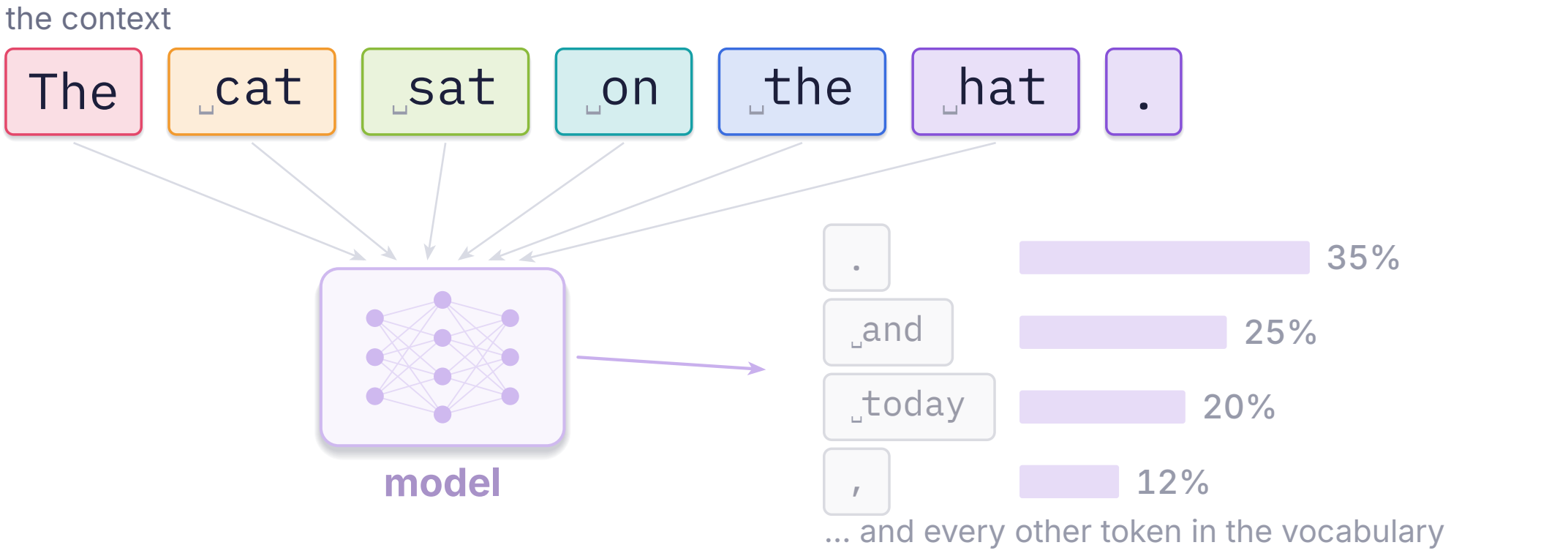
probabilities are illustrative

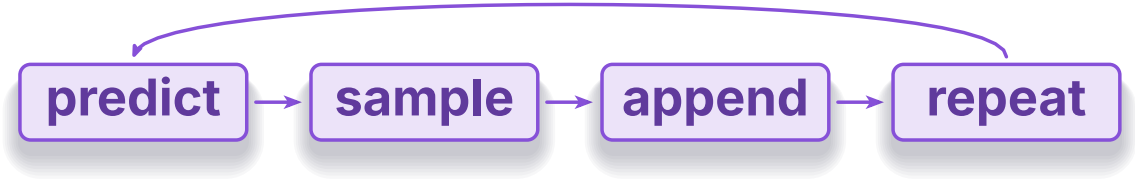
Text is generated one token at a time



probabilities are illustrative

Text is generated one token at a time

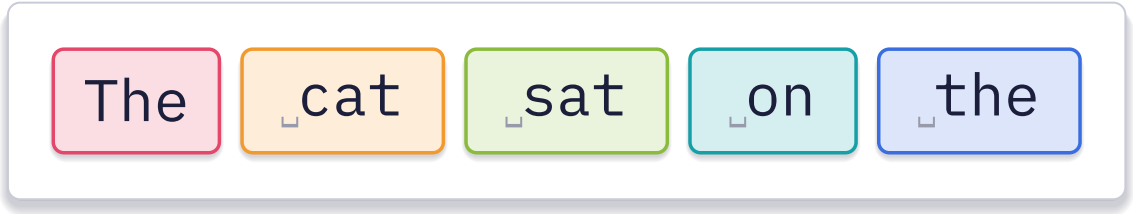




Sampling at every step: the same prompt can give different text.

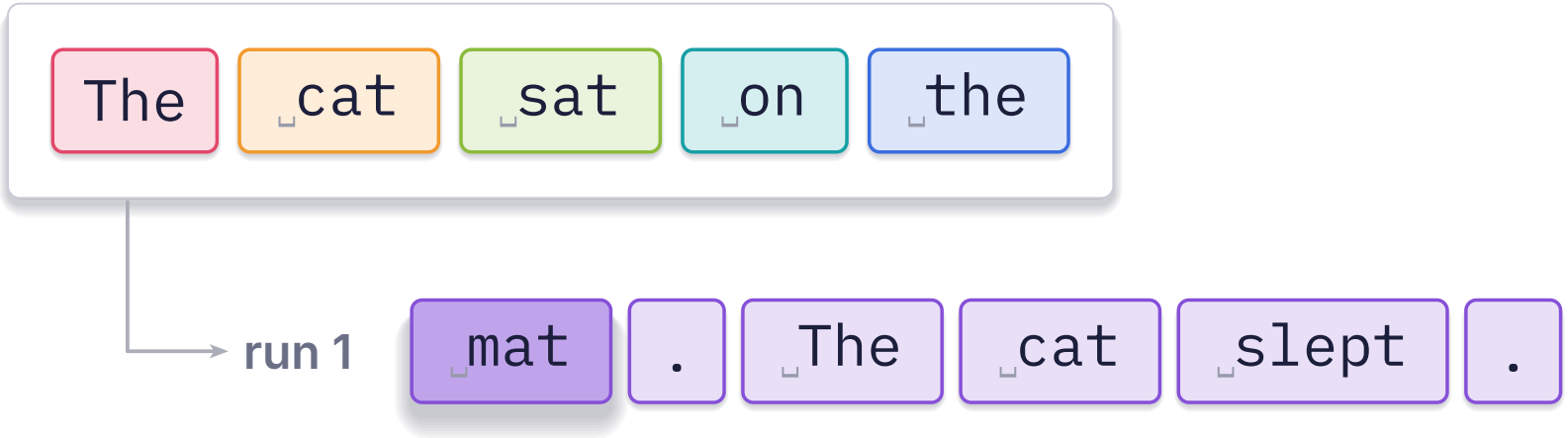
probabilities are illustrative

The same prompt can give different text



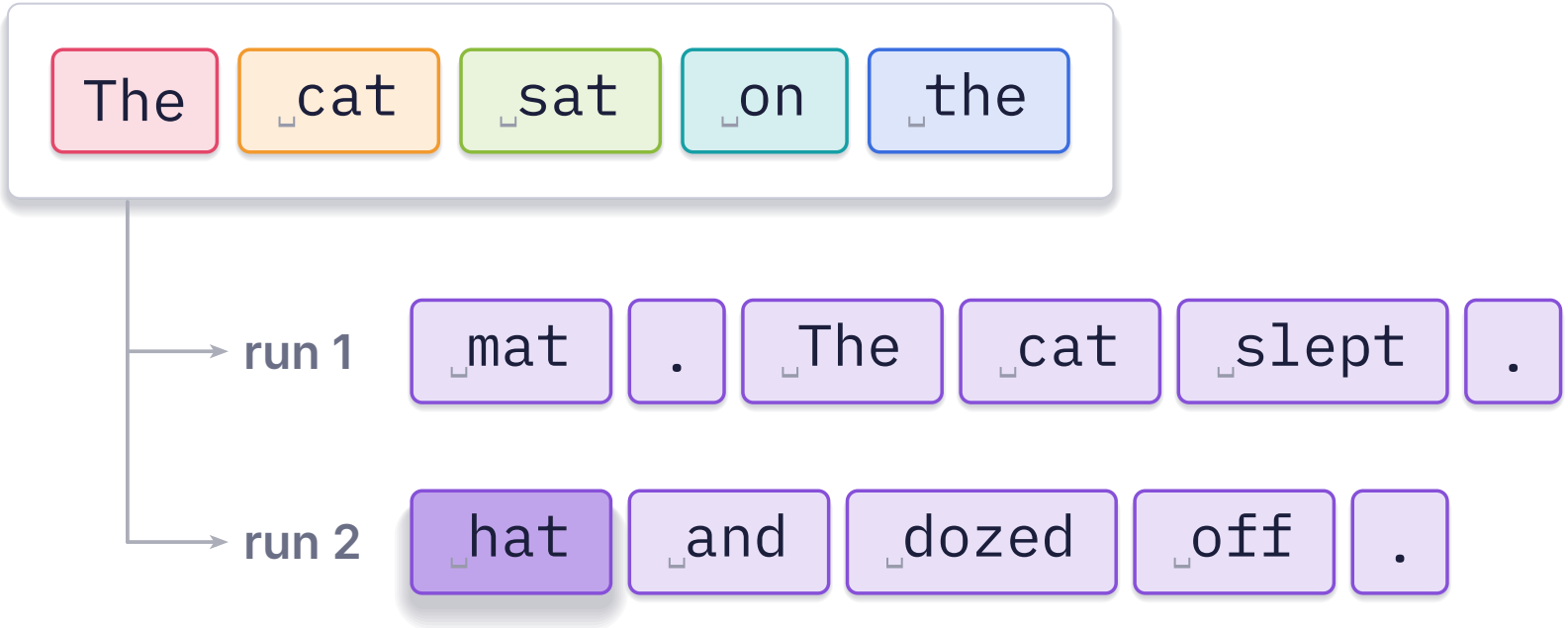
the same prompt, every run

The same prompt can give different text



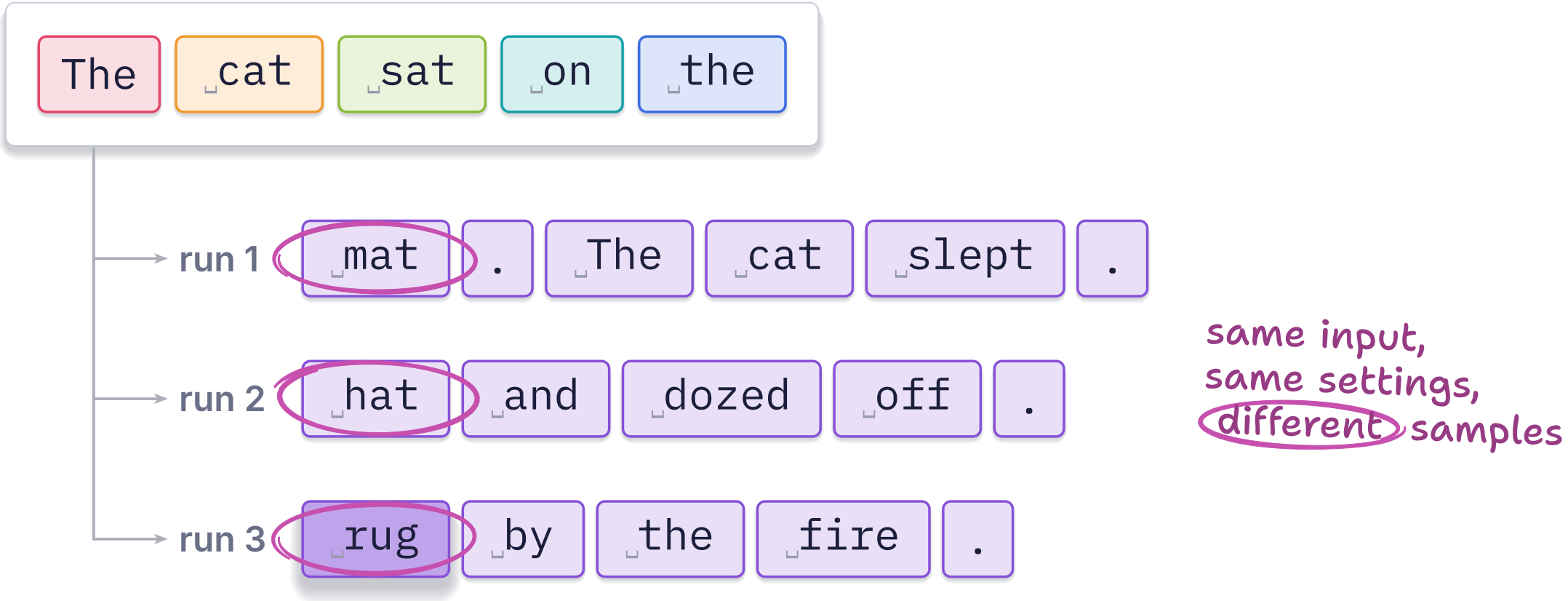
continuations are illustrative

The same prompt can give different text



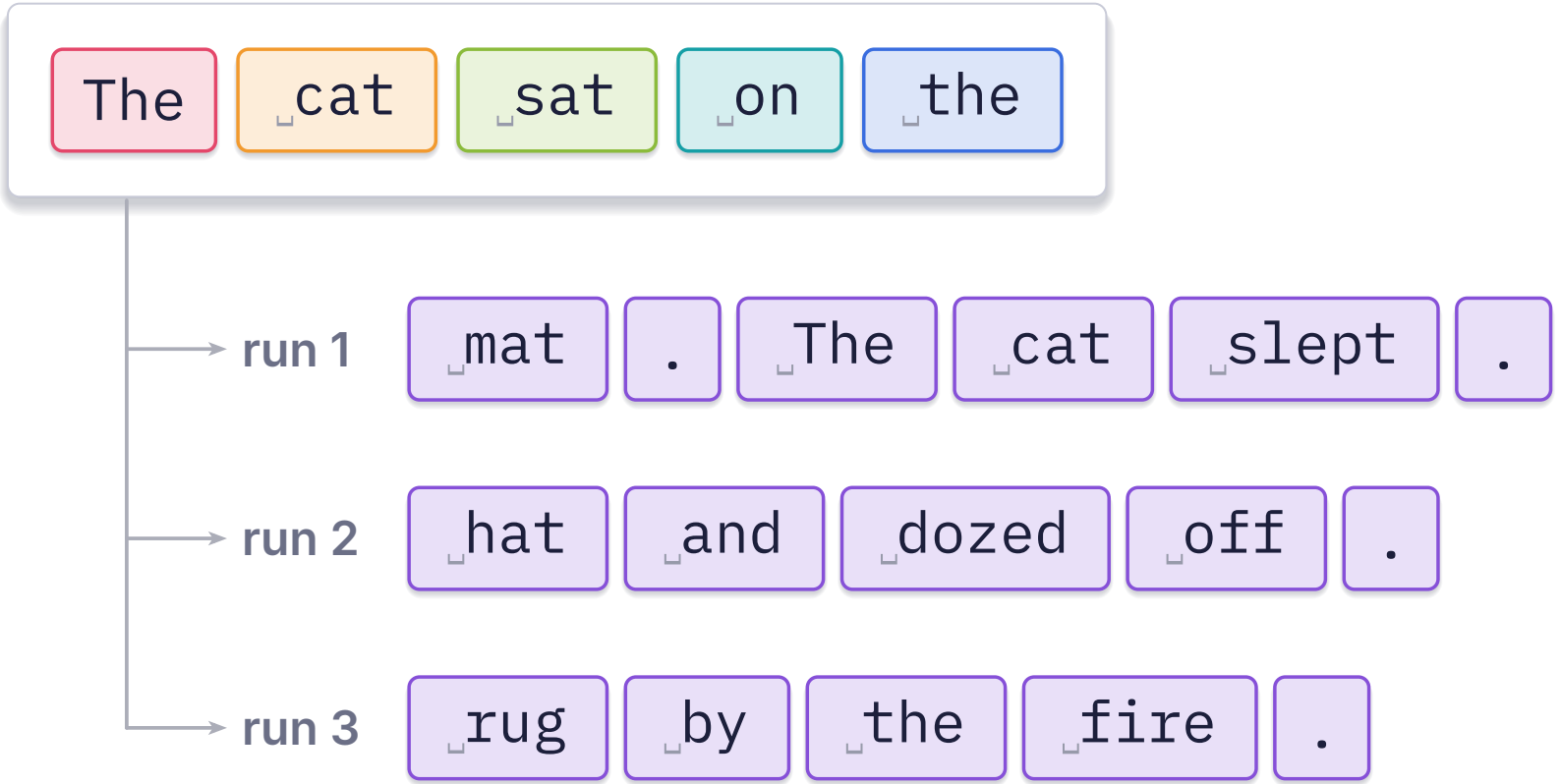
continuations are illustrative

The same prompt can give different text



continuations are illustrative

The same prompt can give different text

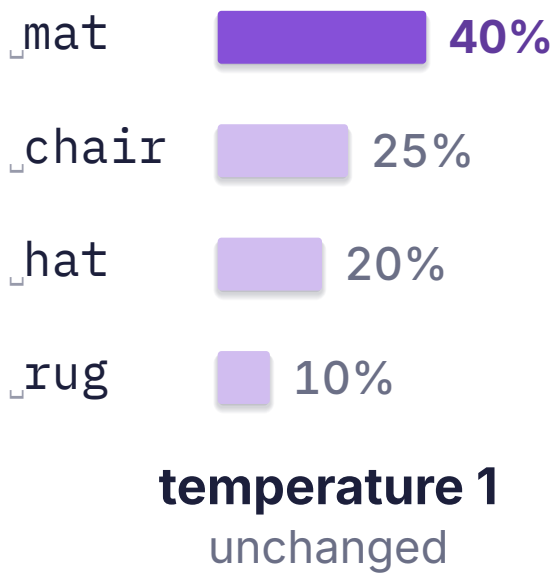


One run is not enough to judge how reliably a prompt works.

Variety helps when you want options: names, explanations, approaches.

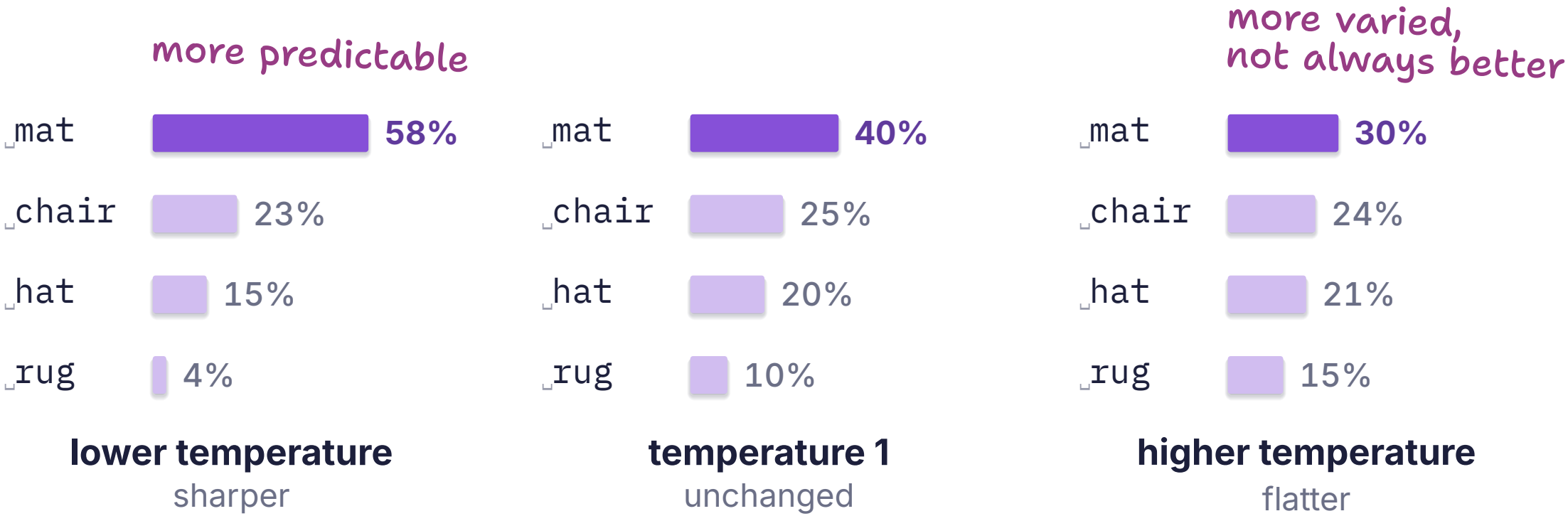
continuations are illustrative

Temperature reshapes the distribution



probabilities are illustrative

Temperature reshapes the distribution

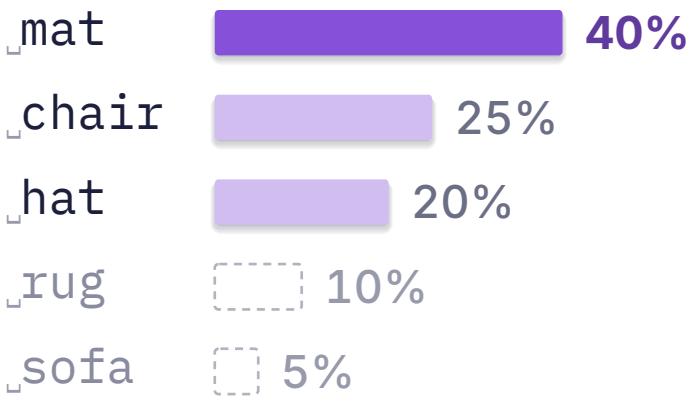


probabilities are illustrative

Top-k and top-p narrow the set we sample from

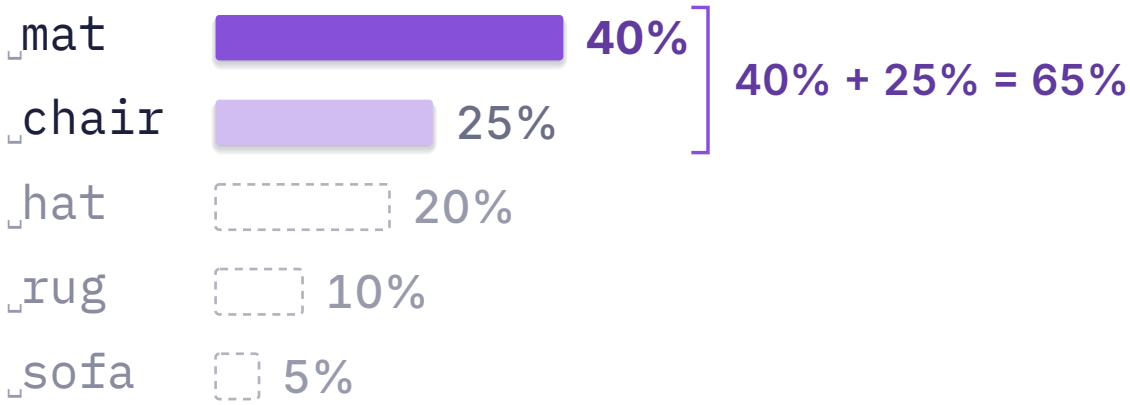
top-k

keep the **3** most likely tokens



top-p

keep tokens until they reach **60%**



---

other controls

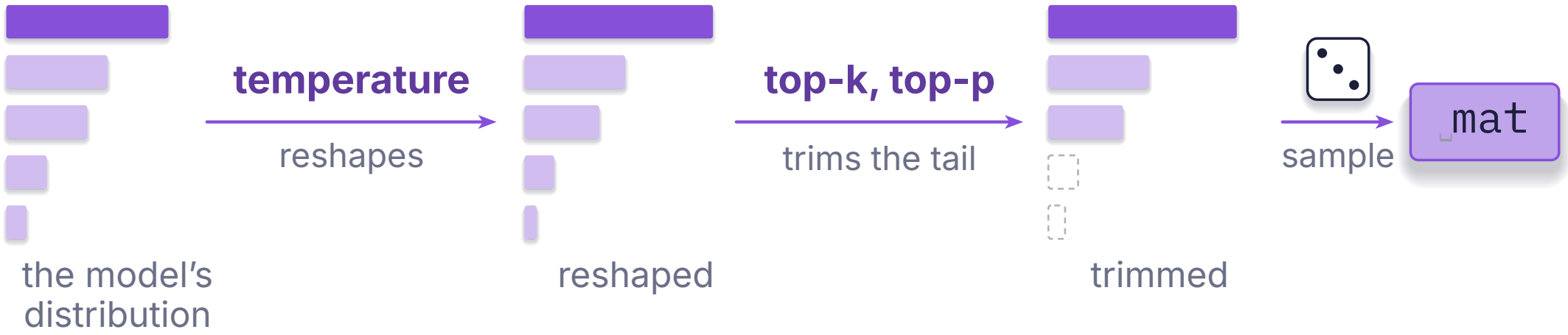
- **a limit on output tokens**
- **a random seed** (in some APIs)

no guarantee: the GPU may add up in another order

```
(0.1+0.2)+0.3 = 0.6000000000000001
0.1+(0.2+0.3) = 0.6
```

probabilities are illustrative

Reshape, trim, then sample one token



probabilities are illustrative

Predicting text well takes real skills

arithmetic

```
If 3 pens cost $6,
one pen costs $ 2
```

needs 6 ÷ 3

Predicting text well takes real skills

arithmetic

```
If 3 pens cost $6,
one pen costs $ 2
```

translation

```
English: cat
French:  chat
```

relates two languages

Predicting text well takes real skills

arithmetic

If 3 pens cost $6,

one pen costs $ 2

translation

English: cat

French: ␣chat

facts

The capital of

Japan is ␣Tokyo

a fact from training

Predicting text well takes real skills

arithmetic

```
If 3 pens cost $6,
one pen costs $ 2
```

translation

```
English: cat
French: ␣chat
```

facts

```
The capital of
Japan is ␣Tokyo
```

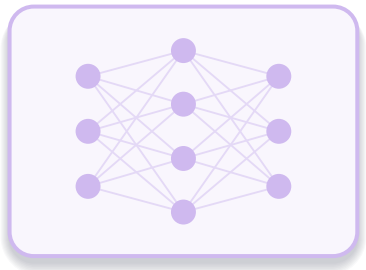
one training goal: predict the next token

each skill can be tested on new problems

A chat is a transcript the model completes

the transcript

`user`	What's the capital of France?
`assistant`	It's Paris.
`user`	And of Japan?
`assistant`	



model

A chat is a transcript the model completes

the transcript

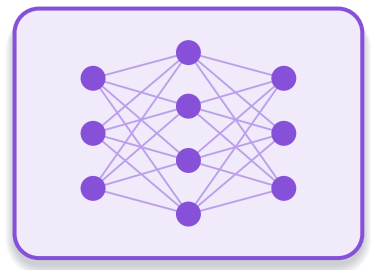
user What's the capital of France?

assistant It's Paris.

user And of Japan?

assistant

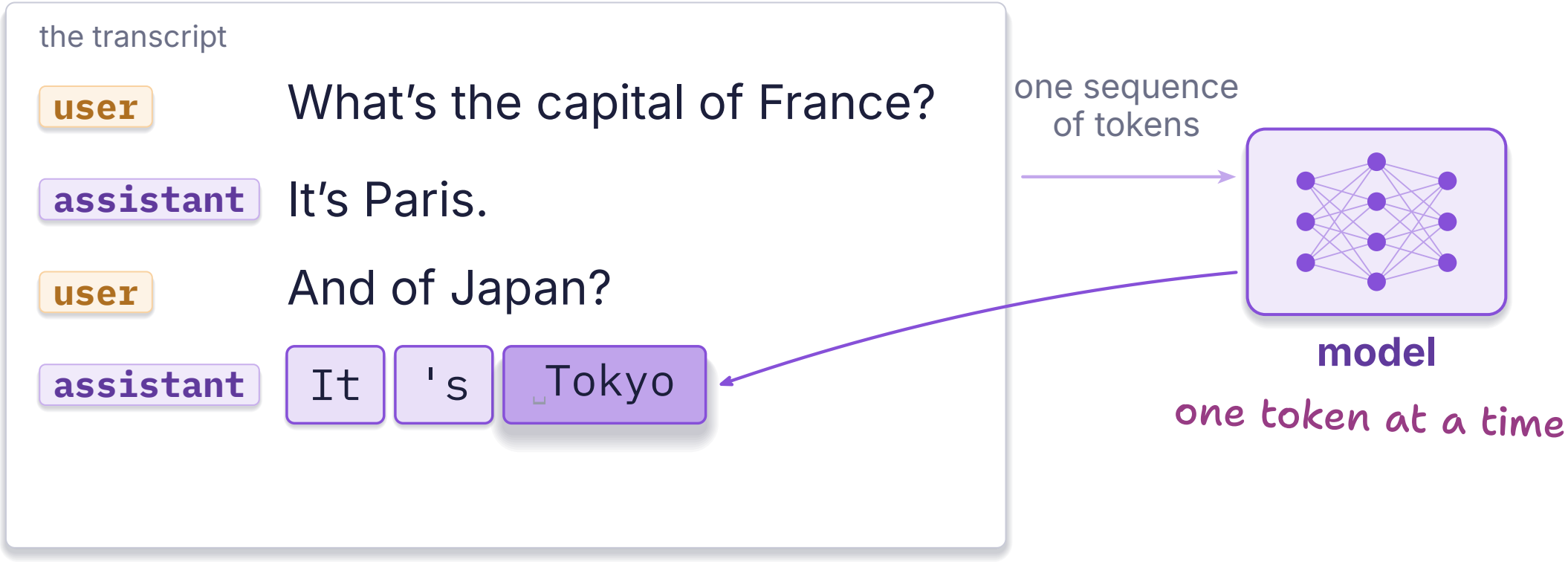
one sequence of tokens



model

the same next-token predictor

A chat is a transcript the model completes

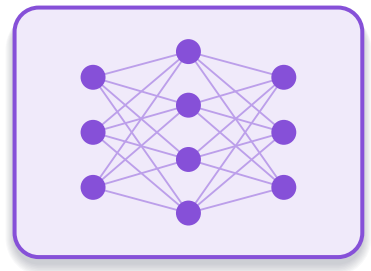


A chat is a transcript the model completes

the transcript

user	What's the capital of France?
assistant	It's Paris.
user	And of Japan?
assistant	`It` `'s` `␣Tokyo` `<end>`
user	*waiting for the user's next message*

one sequence of tokens



model

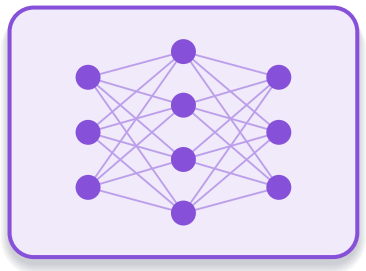
a stop token: the reply ends

A chat is a transcript the model completes

the transcript

user	What's the capital of France?
assistant	It's Paris.
user	And of Japan?
assistant	`It` `'s` `␣Tokyo` `<end>`
user	*waiting for the user's next message*

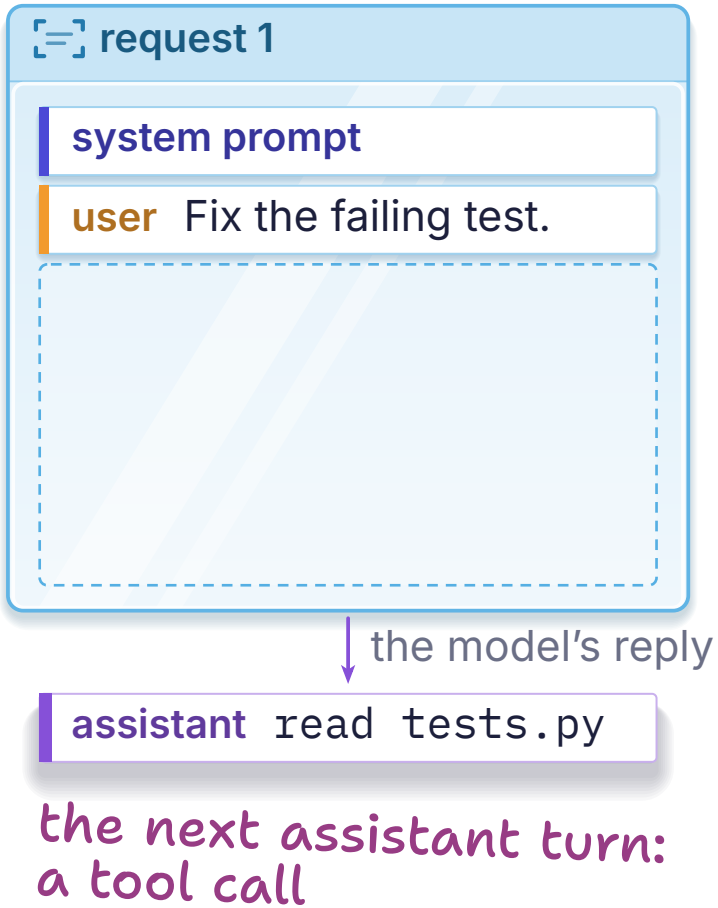
one sequence of tokens →



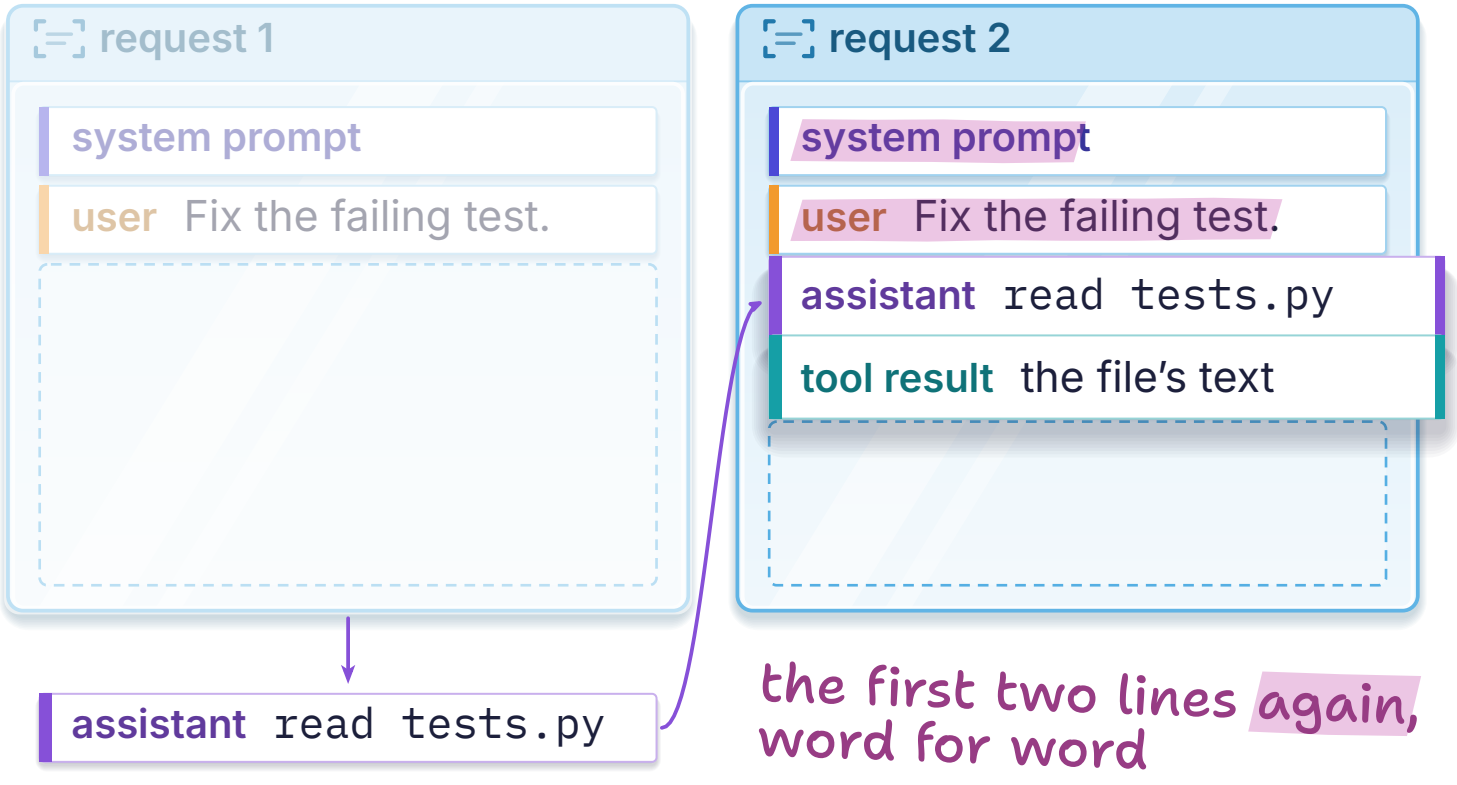
model

Completion over a transcript with role markers.

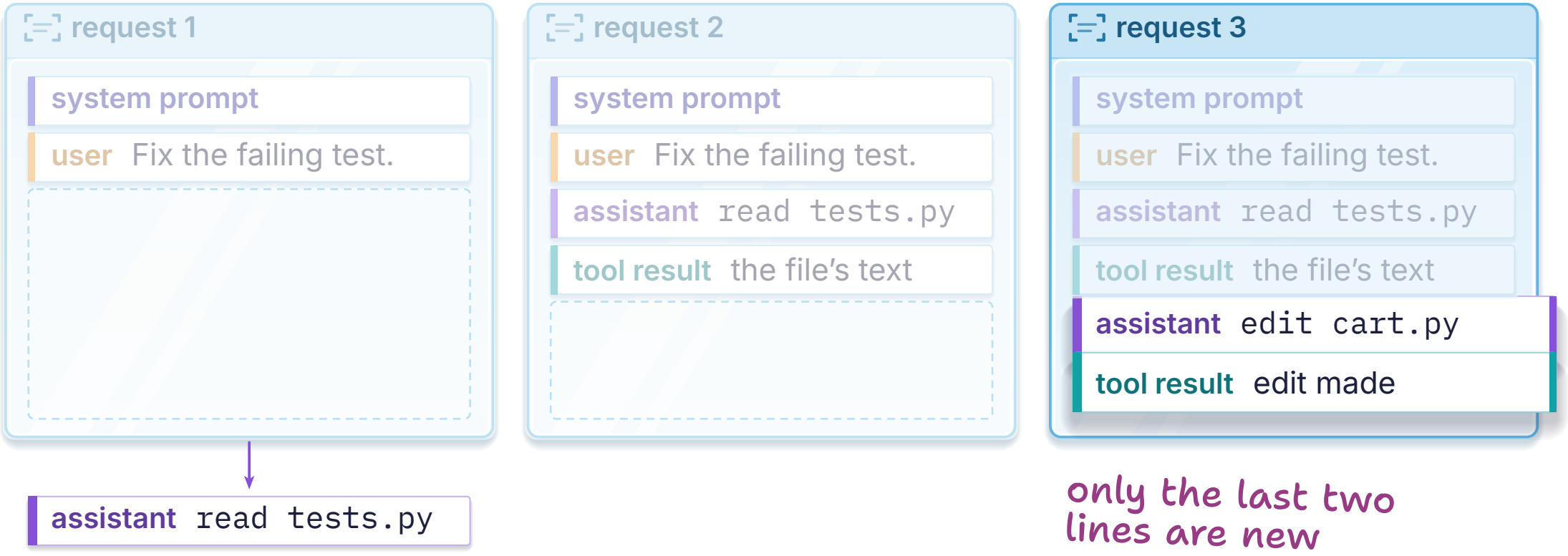
Every request carries the whole conversation



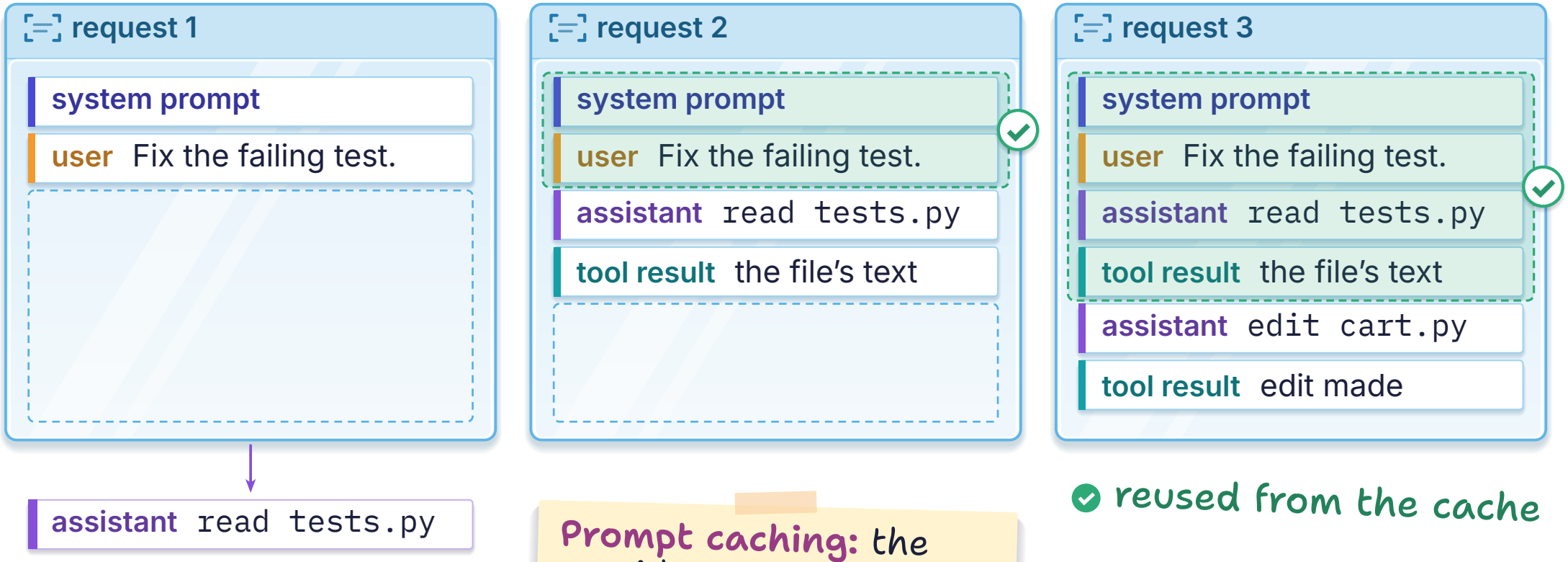
Every request carries the whole conversation



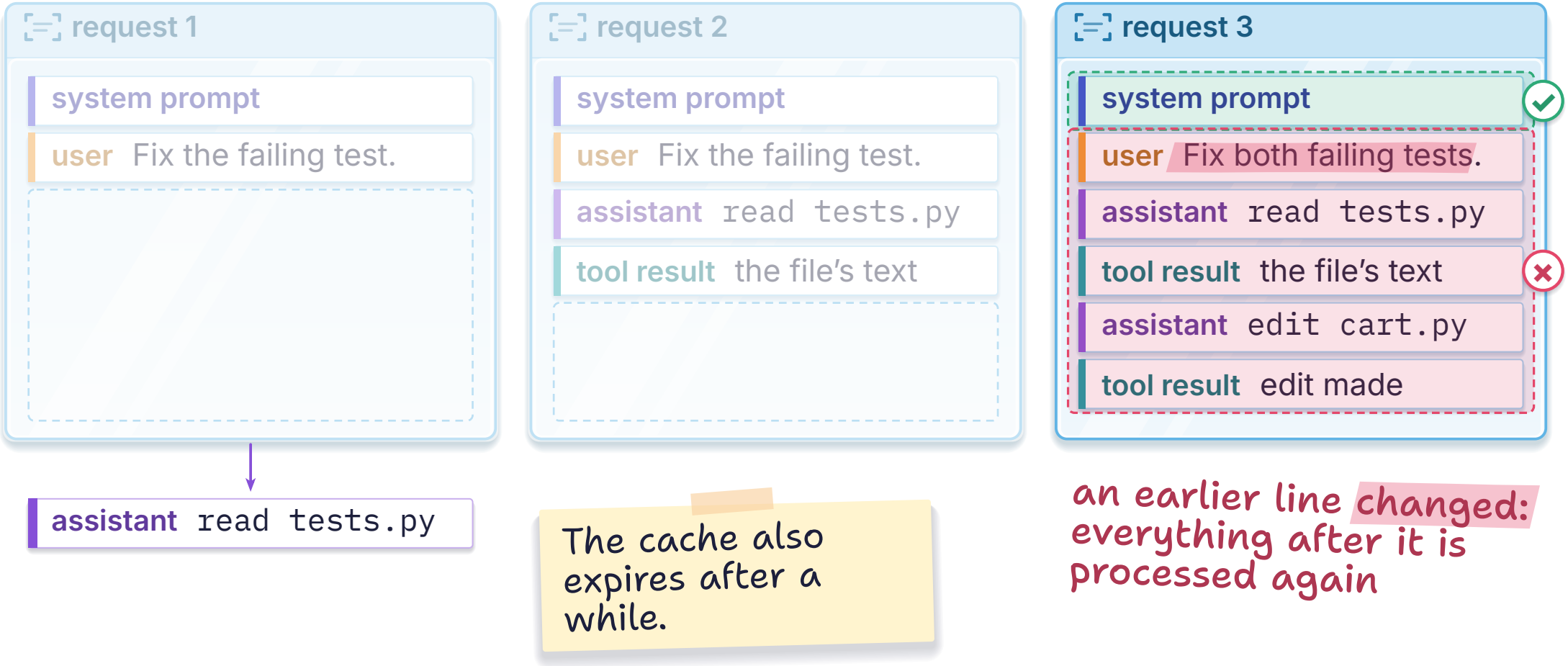
Every request carries the whole conversation



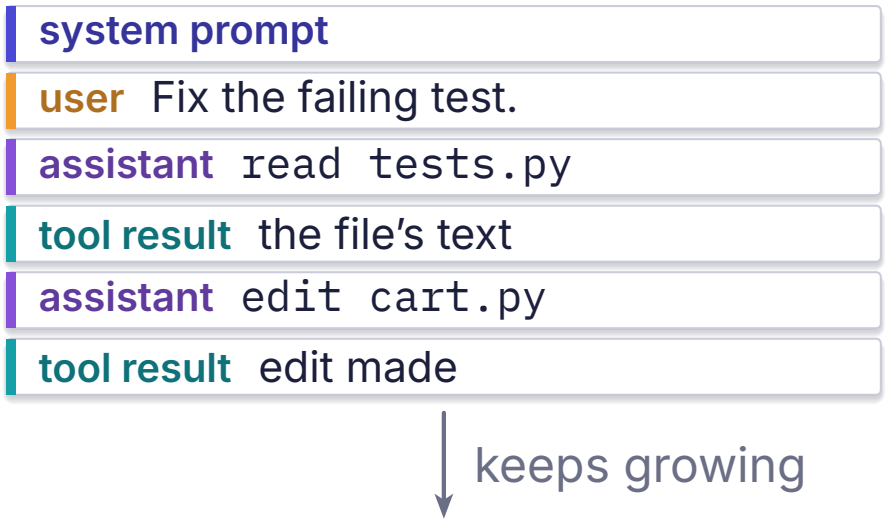
Every request carries the whole conversation



Every request carries the whole conversation



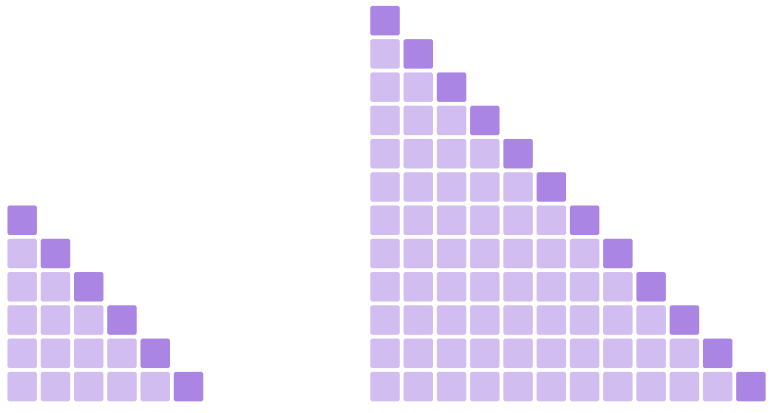
The model sees a fixed number of tokens at once



The model sees a fixed number of tokens at once



each token can use every token before it

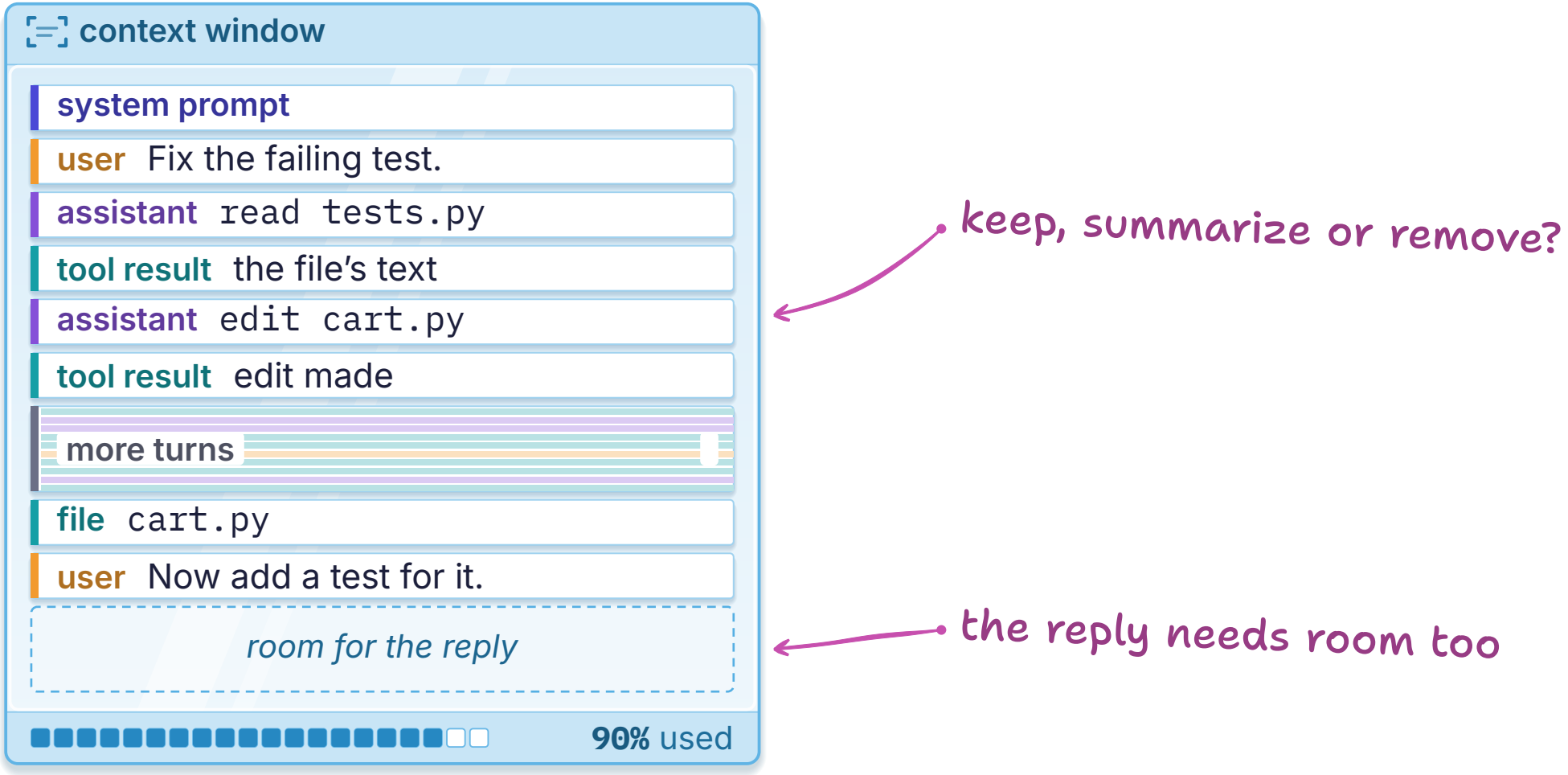


6 tokens
21 comparisons

12 tokens
78 comparisons

twice the tokens,
nearly four times the comparisons

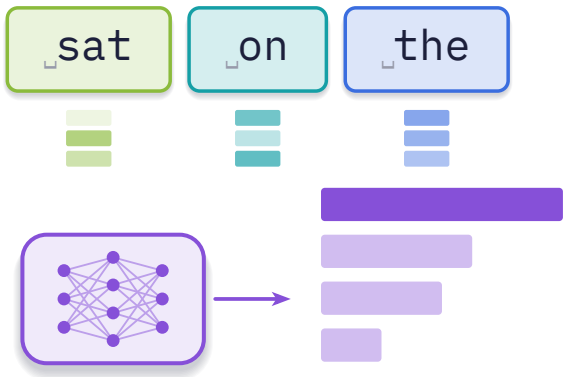
The model sees a fixed number of tokens at once



Recap: from a prediction to a chatting model

Predict

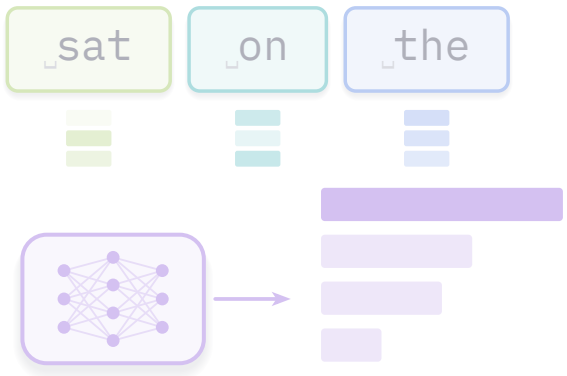
Tokens and learned vectors give a distribution for the next token.



Recap: from a prediction to a chatting model

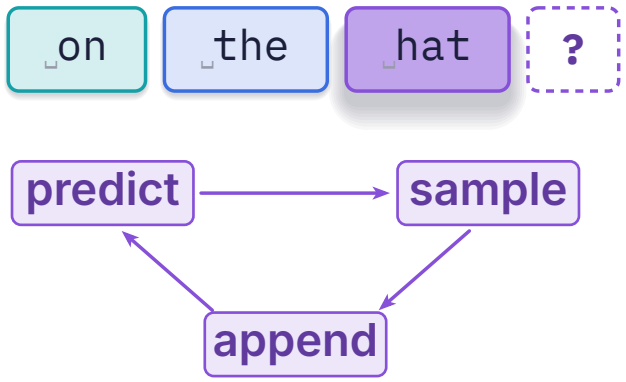
Predict

Tokens and learned vectors give a distribution for the next token.



Generate

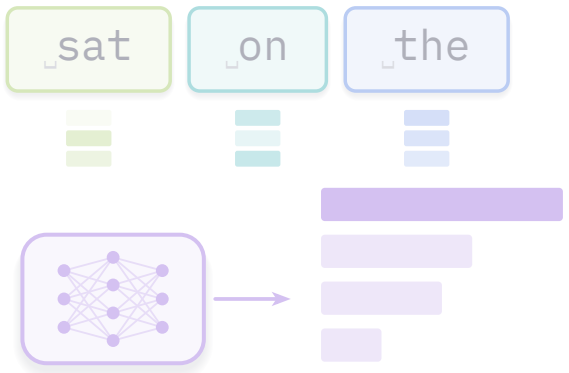
Sample a token, append it, predict again: runs can differ.



Recap: from a prediction to a chatting model

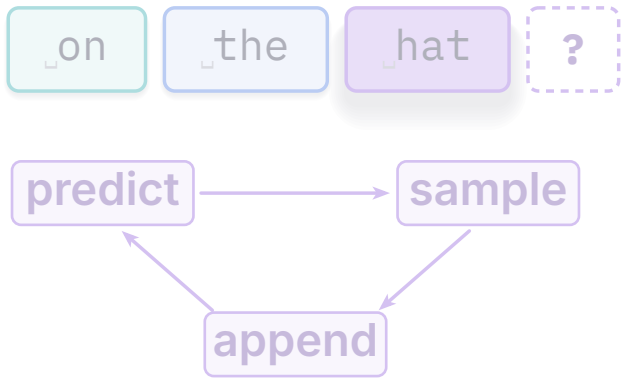
Predict

Tokens and learned vectors give a distribution for the next token.



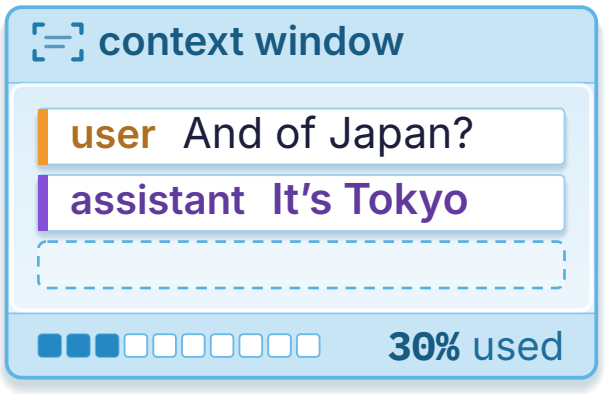
Generate

Sample a token, append it, predict again: runs can differ.



Chat

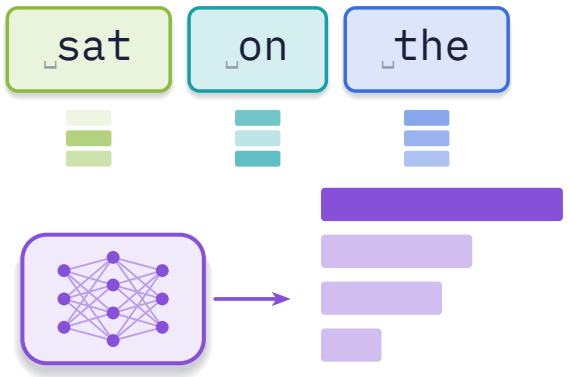
Completion over a transcript with roles, inside a finite window.



Recap: from a prediction to a chatting model

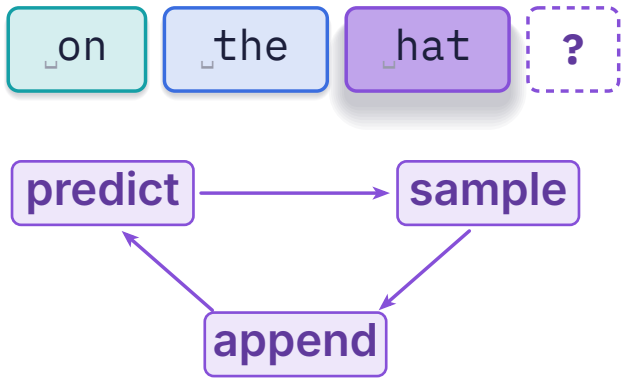
Predict

Tokens and learned vectors give a distribution for the next token.



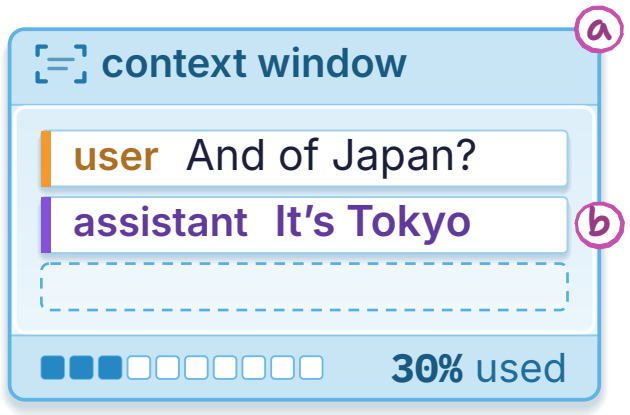
Generate

Sample a token, append it, predict again: runs can differ.



Chat

Completion over a transcript with roles, inside a finite window.



ⓐ What information was in the window?
ⓑ What did the model generate from it?