next-token prediction

artificial intelligence · prof. dr. nicolas meseth

1. tokens
2. one at a time
3. choosing
4. the context
5. the transformer

every answer
of a language
model is written
one token
at a time.



image: ai-generated (gpt-image-2.5-flare)

part 1

tokens

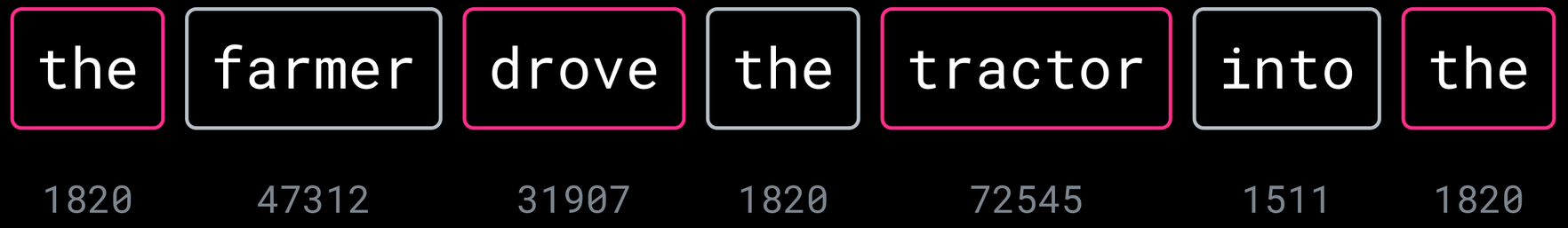
a token is a piece of text



split and numbers for illustration

tokens

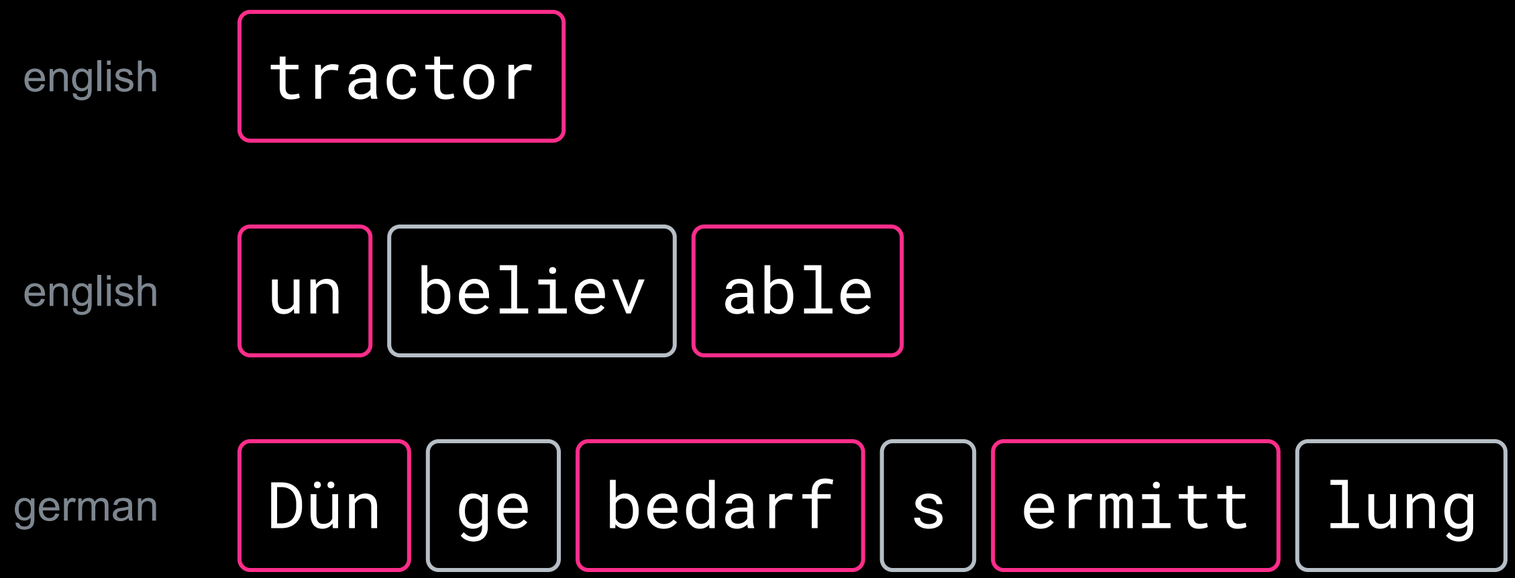
a token is a piece of text



split and numbers for illustration

tokens

rare words fall apart



split for illustration · try it yourself: tiktokenizer.vercel.app

tokens

rare words fall apart



split for illustration · try it yourself: tiktokenizer.vercel.app

tokens

262,000

tokens in the vocabulary of Gemma 4 E2B, and everything the model writes is put together from them.

source: gemma 4 model card

part 2

one at a time

in one step, what does the model compute?

the next word

the whole answer

a probability per token

a search on the internet

lifi.uber.space/q

in one step, what does the model compute?

- the next word
- the whole answer
- a probability per token
- a search on the internet

lifi.uber.space/q

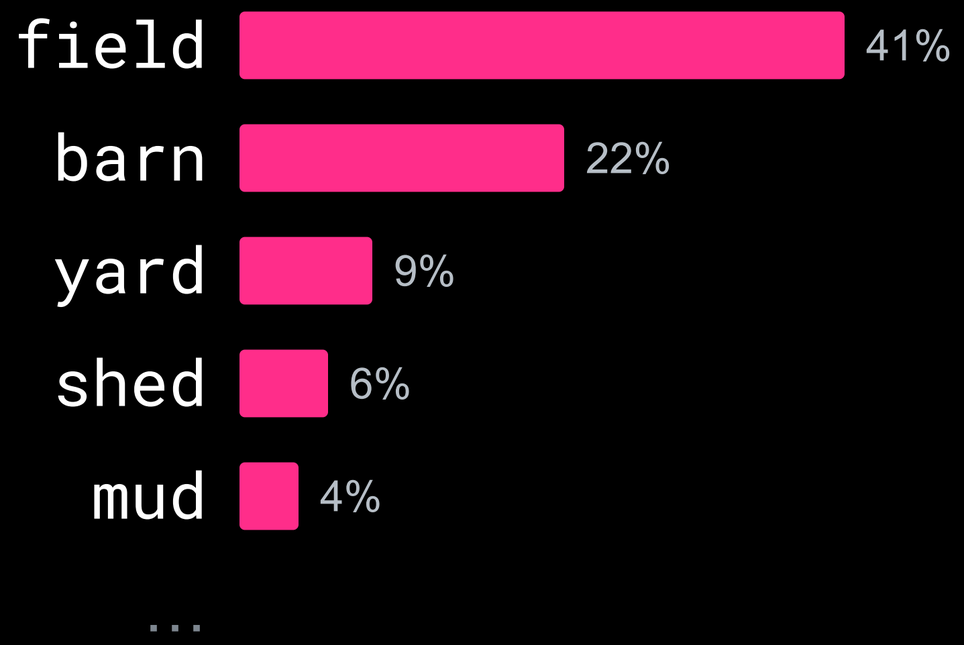
in one step, what does the model compute?

- the next word
- the whole answer
- a probability per token
- a search on the internet

lifi.uber.space/q

the field, the barn or the yard?

the farmer drove the tractor into the …

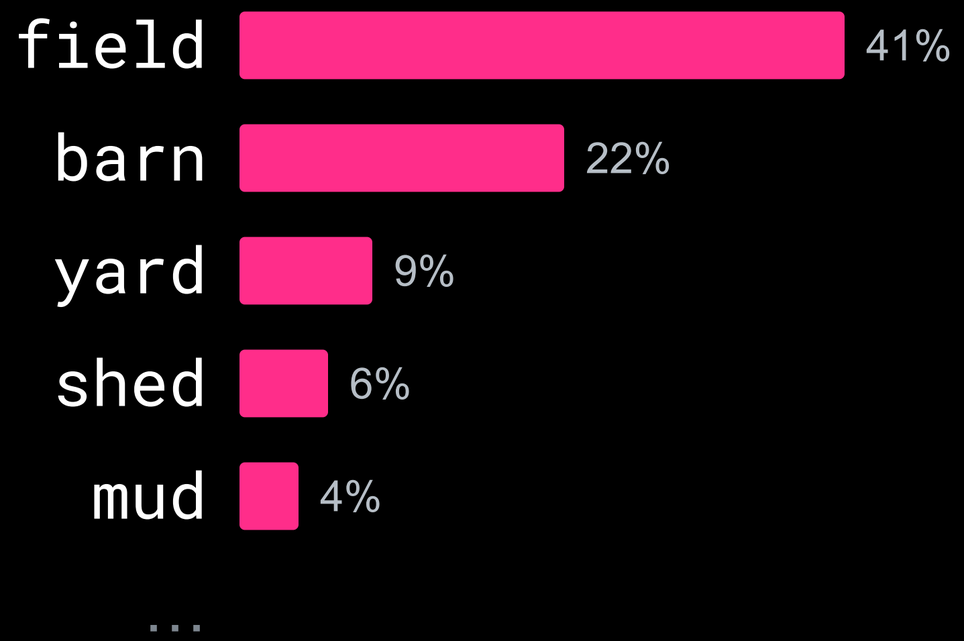


probabilities for illustration · try it with a real model: alonsosilva-nexttokenprediction.hf.space

one at a time

the field, the barn or the yard?

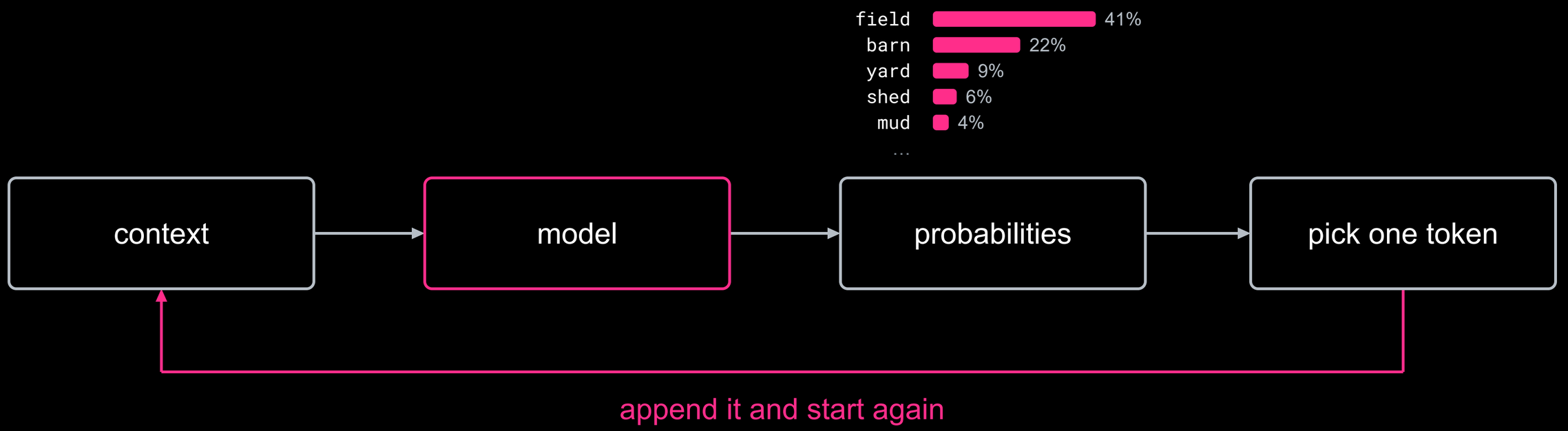
the farmer drove the tractor into the …



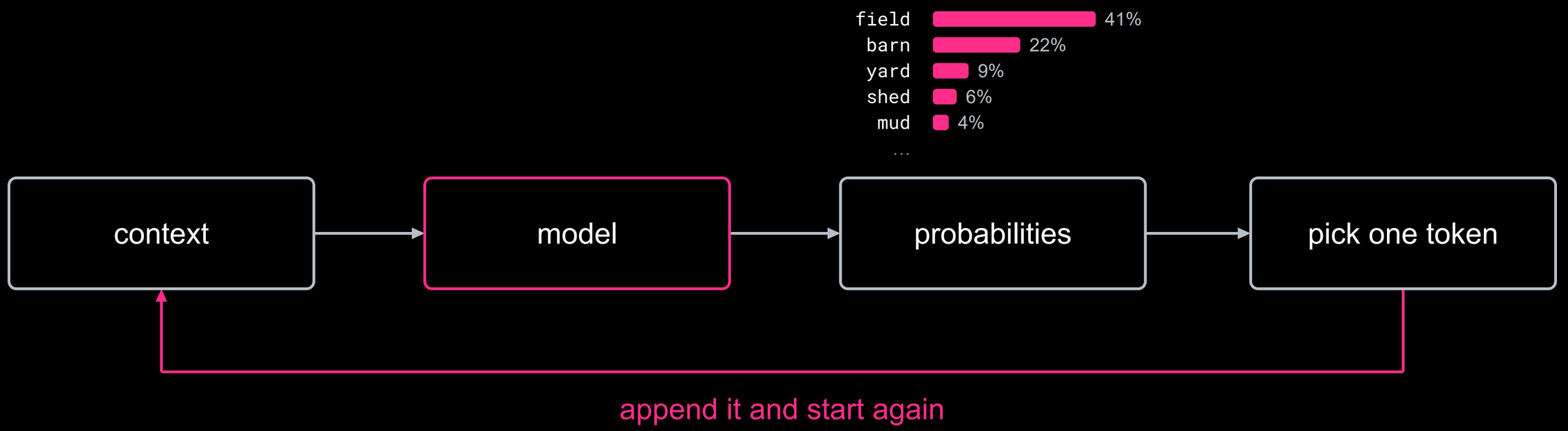
262,000 tokens in the vocabulary, most of them almost zero

probabilities for illustration · try it with a real model: alonsosilva-nexttokenprediction.hf.space

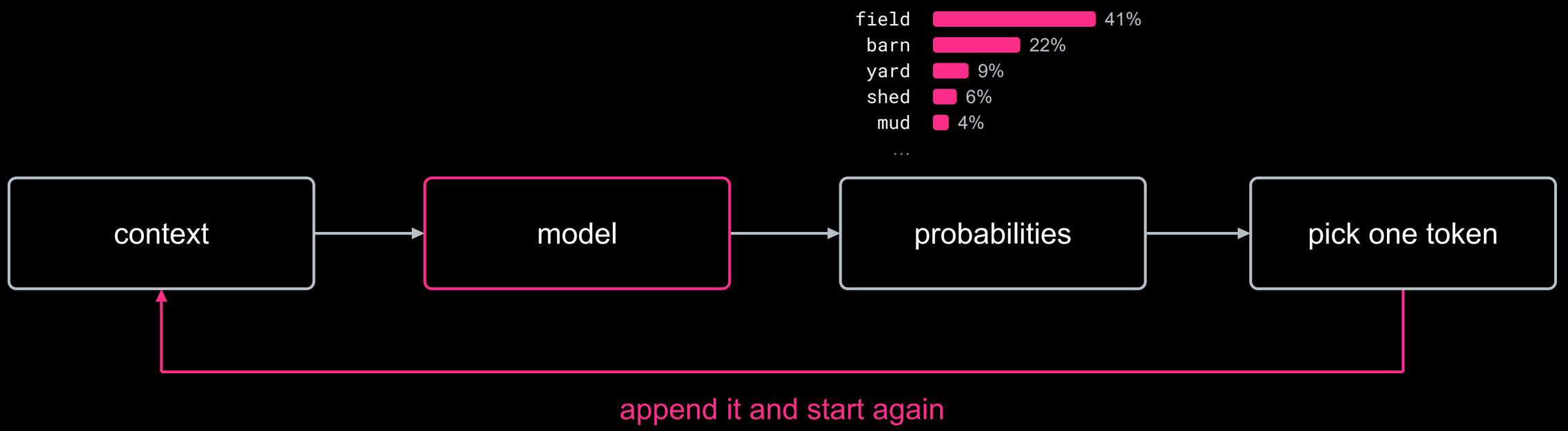
one at a time



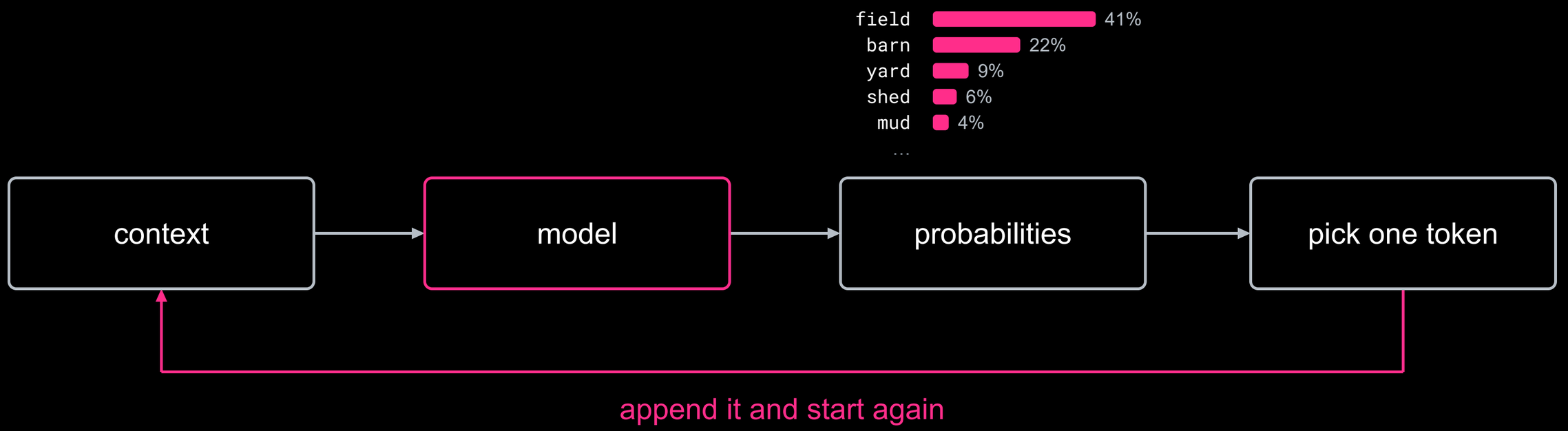
the farmer drove the tractor into the



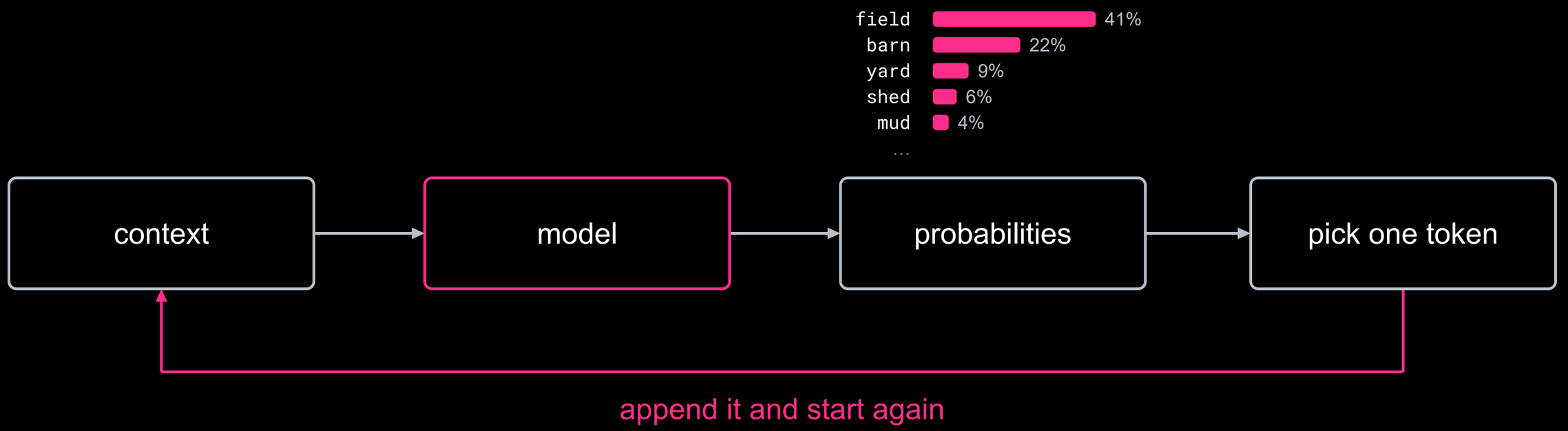

```
the farmer drove the tractor into the field
```



the farmer drove the tractor into the field,



the farmer drove the tractor into the field, where



`the farmer drove the tractor into the field, where`

? when does it stop?

the model writes the most likely continuation, which is not always the true one.

part 3

choosing

ask the same question twice: why do you get two different answers?

the model learned

tokens are drawn at random

another model answered

it searched again



lifi.uber.space/q

ask the same question twice: why do you get two different answers?

- the model learned
- tokens are drawn at random
- another model answered
- it searched again

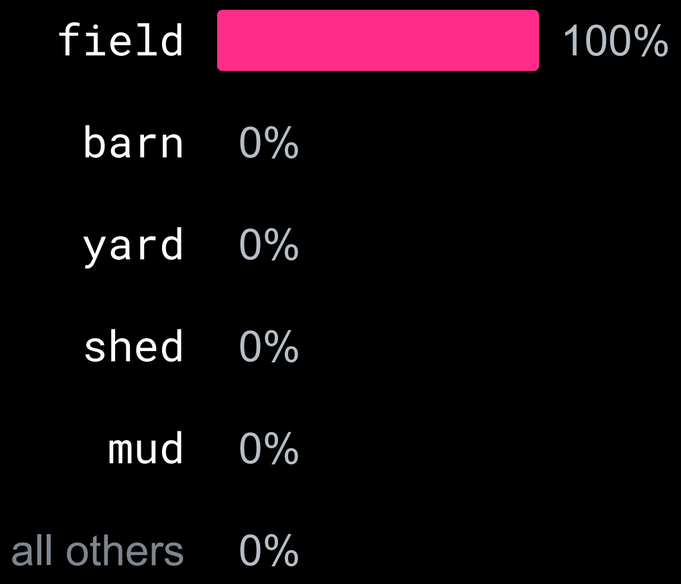
lifi.uber.space/q

ask the same question twice: why do you get two different answers?

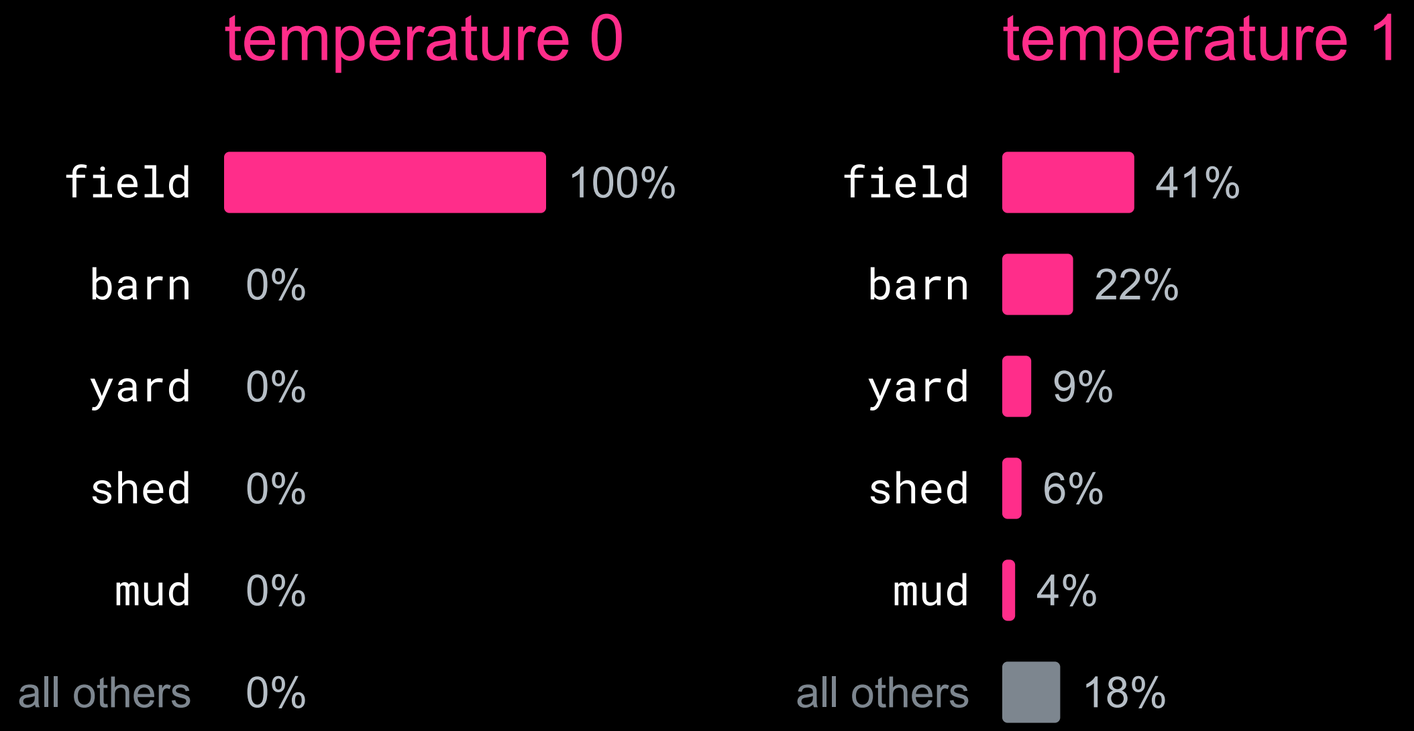
- the model learned
- tokens are drawn at random
- another model answered
- it searched again

lifi.uber.space/q

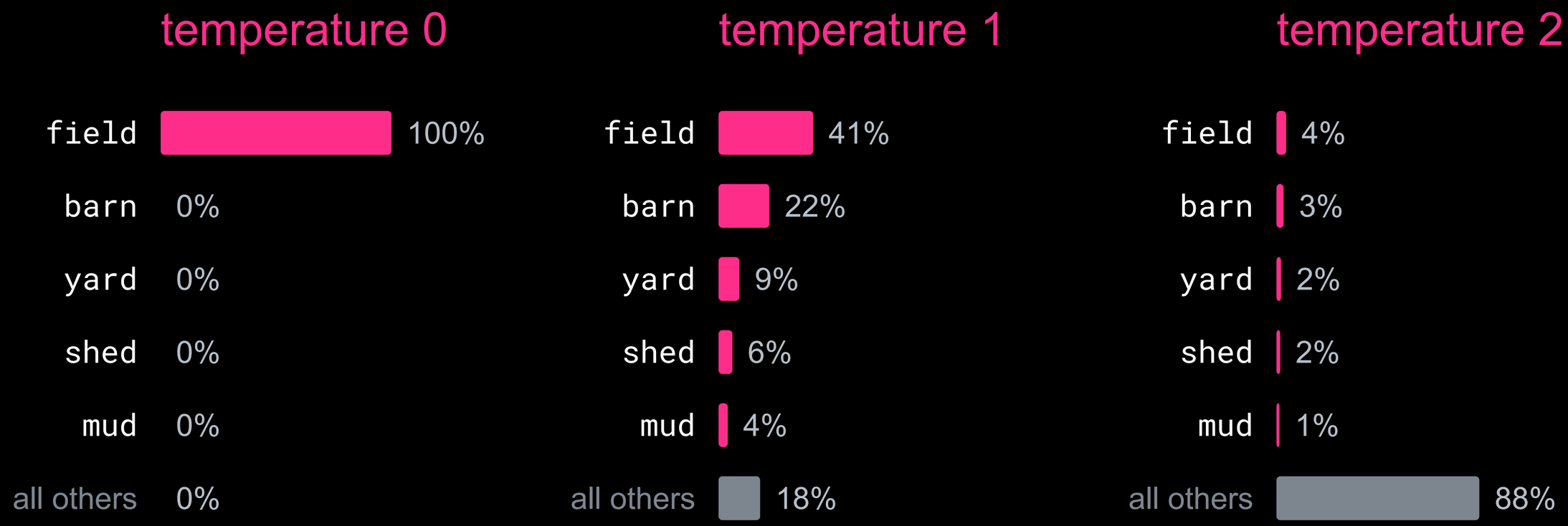
temperature 0



probabilities for illustration, temperature calculated · try it live: poloclub.github.io/transformer-explainer

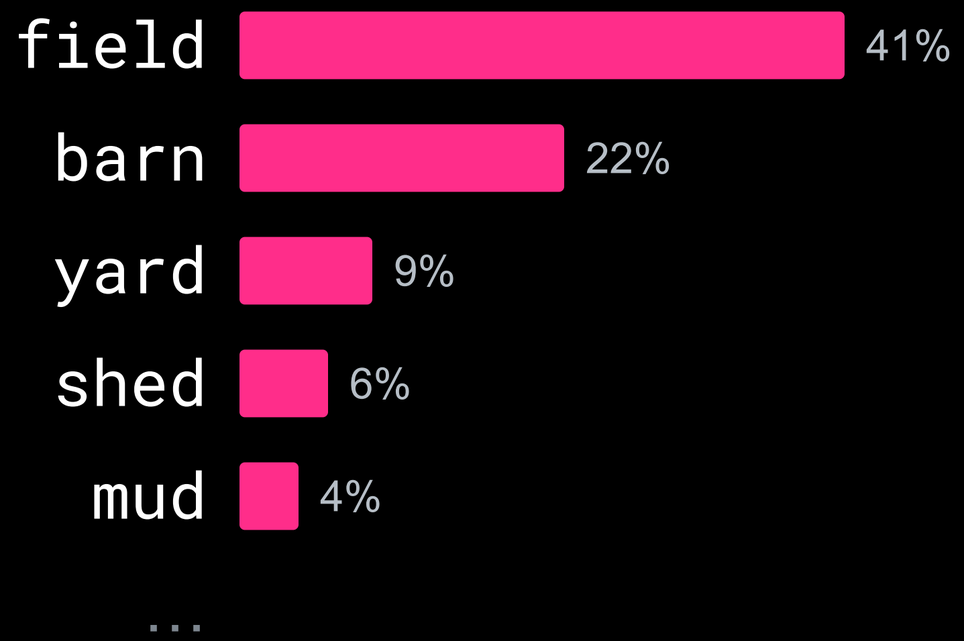


probabilities for illustration, temperature calculated · try it live: poloclub.github.io/transformer-explainer



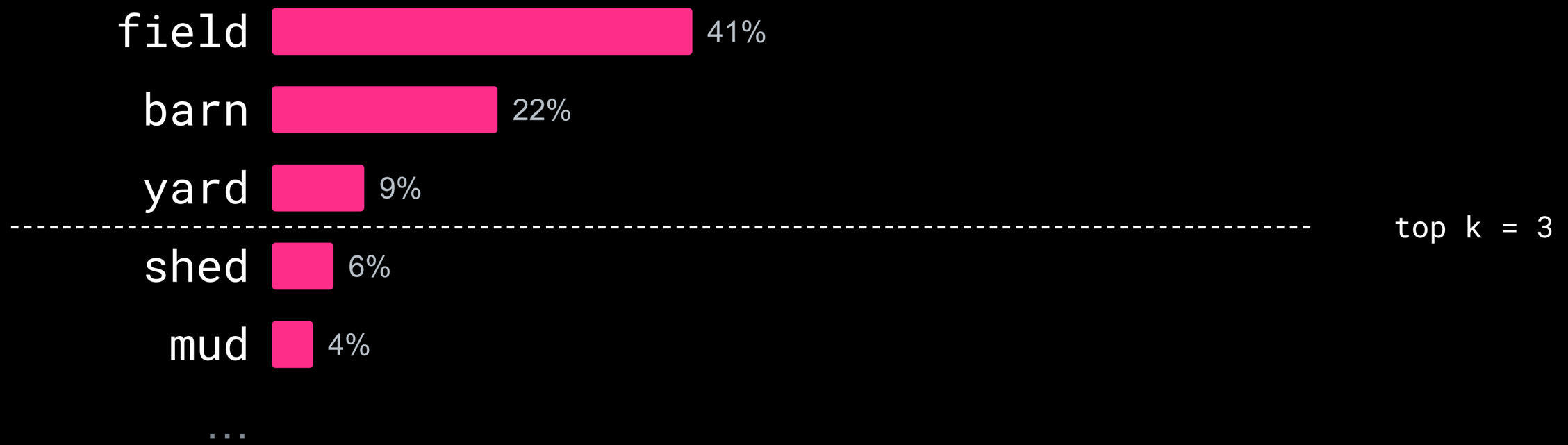
probabilities for illustration, temperature calculated · try it live: poloclub.github.io/transformer-explainer

the farmer drove the tractor into the …



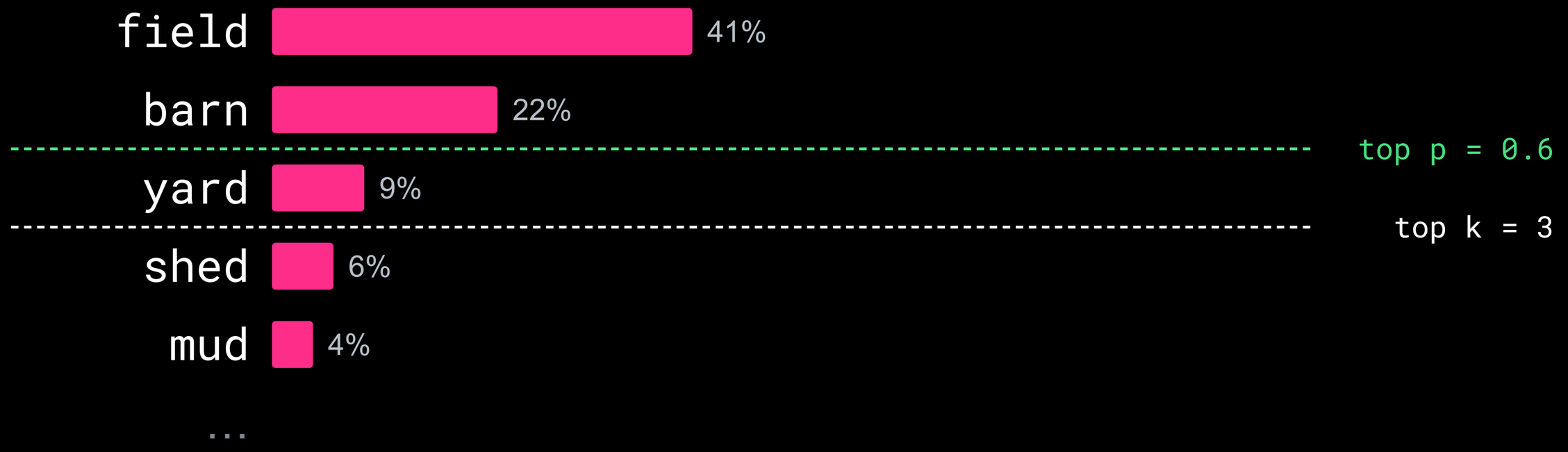
probabilities for illustration · try it live: poloclub.github.io/transformer-explainer

the farmer drove the tractor into the …



probabilities for illustration · try it live: poloclub.github.io/transformer-explainer

the farmer drove the tractor into the …



probabilities for illustration · try it live: poloclub.github.io/transformer-explainer

ask the same question three times at temperature 0.

in LM Studio, switch to developer mode to see all settings.

ask the same question three times at temperature 0.

ask it three times at temperature 1.

in LM Studio, switch to developer mode to see all settings.

ask the same question three times at temperature 0.

ask it three times at temperature 1.

switch off the filters: `top k = 0`, `top p = 1`, `min p = 0`.

in LM Studio, switch to developer mode to see all settings.

ask the same question three times at temperature 0.

ask it three times at temperature 1.

switch off the filters: `top k = 0`, `top p = 1`, `min p = 0`.

raise the temperature step by step until the answer falls apart.

in LM Studio, switch to developer mode to see all settings.

part 4

the context

everything it sees

system prompt

the context

everything it sees

the conversation so far

system prompt

everything it sees

your document

the conversation so far

system prompt

everything it sees

the answer so far

your document

the conversation so far

system prompt

everything it sees

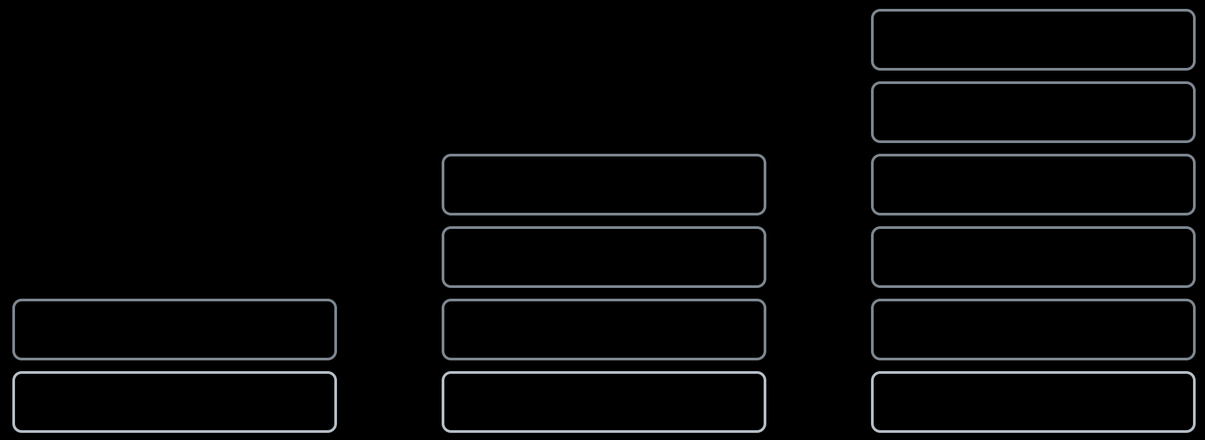
the answer so far

your document

the conversation so far

system prompt

the program sends everything again,
because the model remembers nothing



message 1 message 2 message 3

how many tokens fit into the context of Gemma 4 E2B?

4,096

32,768

131,072

no limit

lifi.uber.space/q

how many tokens fit into the context of Gemma 4 E2B?

- 4,096
- 32,768
- 131,072
- no limit

lifi.uber.space/q

how many tokens fit into the context of Gemma 4 E2B?

- 4,096
- 32,768
- 131,072
- no limit

lifi.uber.space/q

the model was trained on texts up to a certain length.

the model was trained on texts up to a certain length.

every new token is related to every token before it.

the model was trained on texts up to a certain length.

every new token is related to every token before it.

every token in the context needs memory.

the model was trained on texts up to a certain length.

every new token is related to every token before it.

every token in the context needs memory.

when the context is full, something has to go.

stop at limit

the answer ends when the window is full.

stop at limit	the answer ends when the window is full.
rolling window	the oldest messages are dropped.

stop at limit	the answer ends when the window is full.
rolling window	the oldest messages are dropped.
truncate middle	the beginning and the latest messages stay.

stop at limit	the answer ends when the window is full.
rolling window	the oldest messages are dropped.
truncate middle	the beginning and the latest messages stay.
summarise	the old part is replaced by a short summary.

stop at limit	the answer ends when the window is full.
rolling window	the oldest messages are dropped.
truncate middle	the beginning and the latest messages stay.
summarise	the old part is replaced by a short summary.

whatever is dropped no longer exists for the model.

part 5

the transformer

the tractor or the gate?

the tractor did not fit through the gate because it was too wide

to predict the next token, a transformer looks at all tokens of the context at once

a. vaswani et al.: attention is all you need, 2017

the tractor or the gate?



the tractor did not fit through the gate because it was too wide

to predict the next token, a transformer looks at all tokens of the context at once

a. vaswani et al.: attention is all you need, 2017

the tractor or the gate?

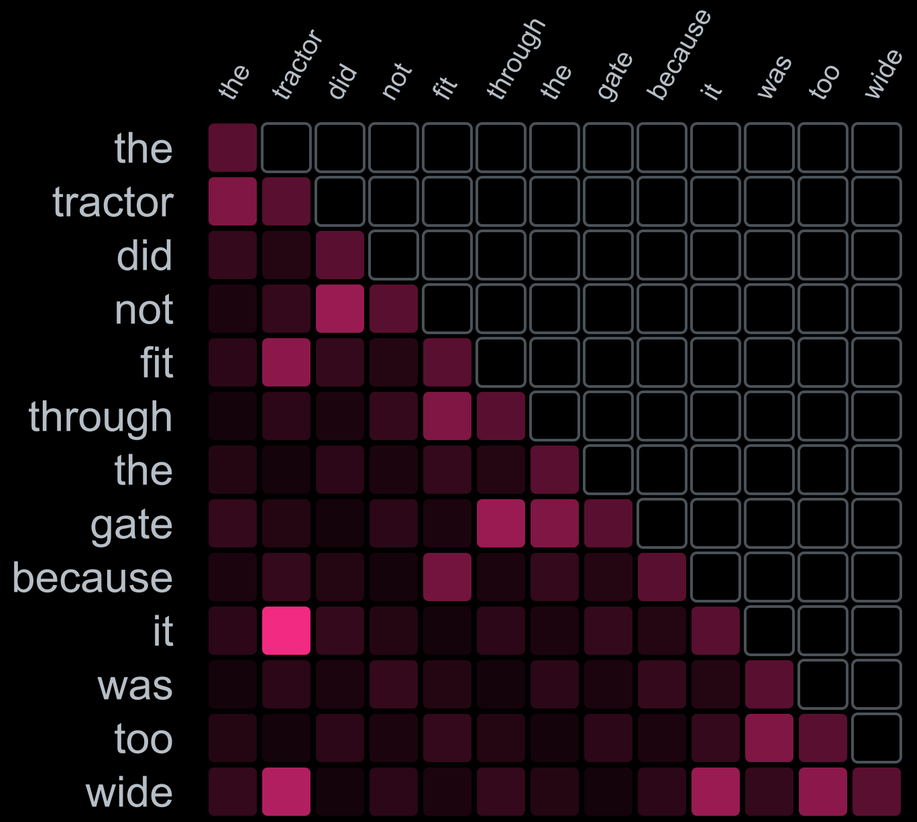


the tractor did not fit through the gate because it was too narrow

to predict the next token, a transformer looks at all tokens of the context at once

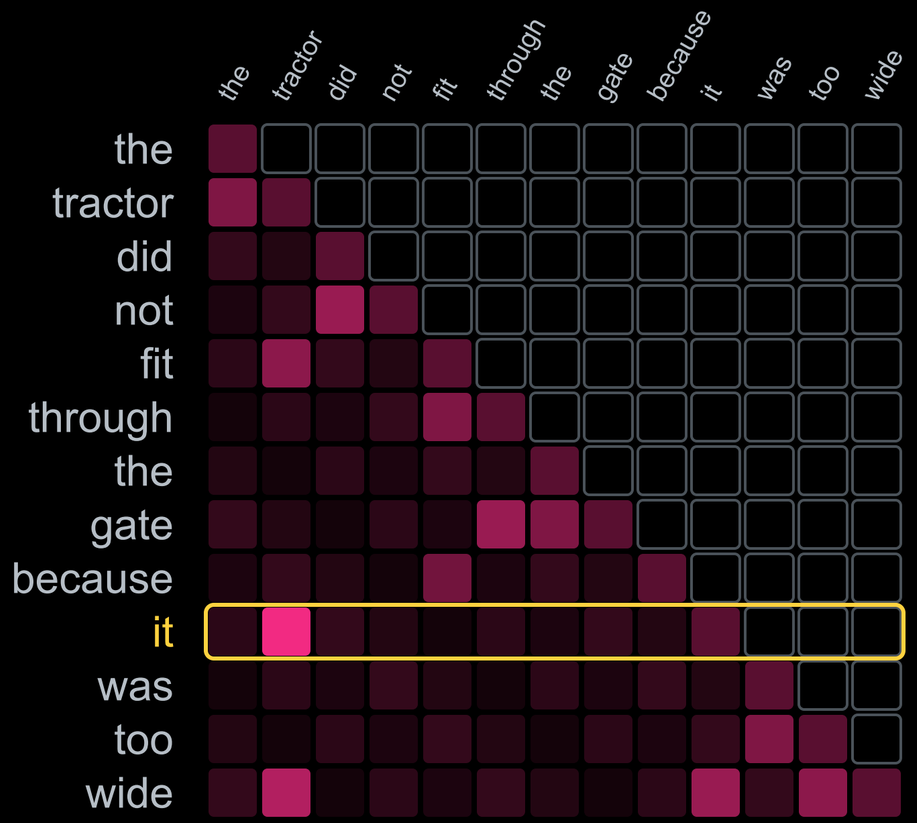
a. vaswani et al.: attention is all you need, 2017

every token with every token



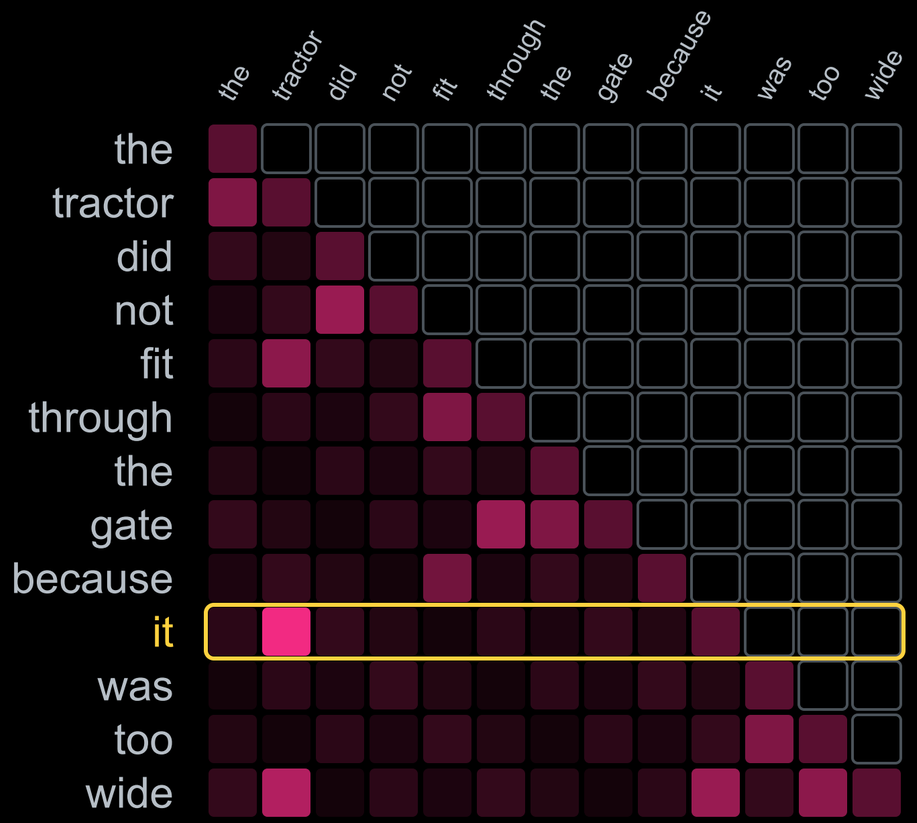
words instead of tokens and strengths for illustration

every token with every token



words instead of tokens and strengths for illustration

every token with every token



tokens

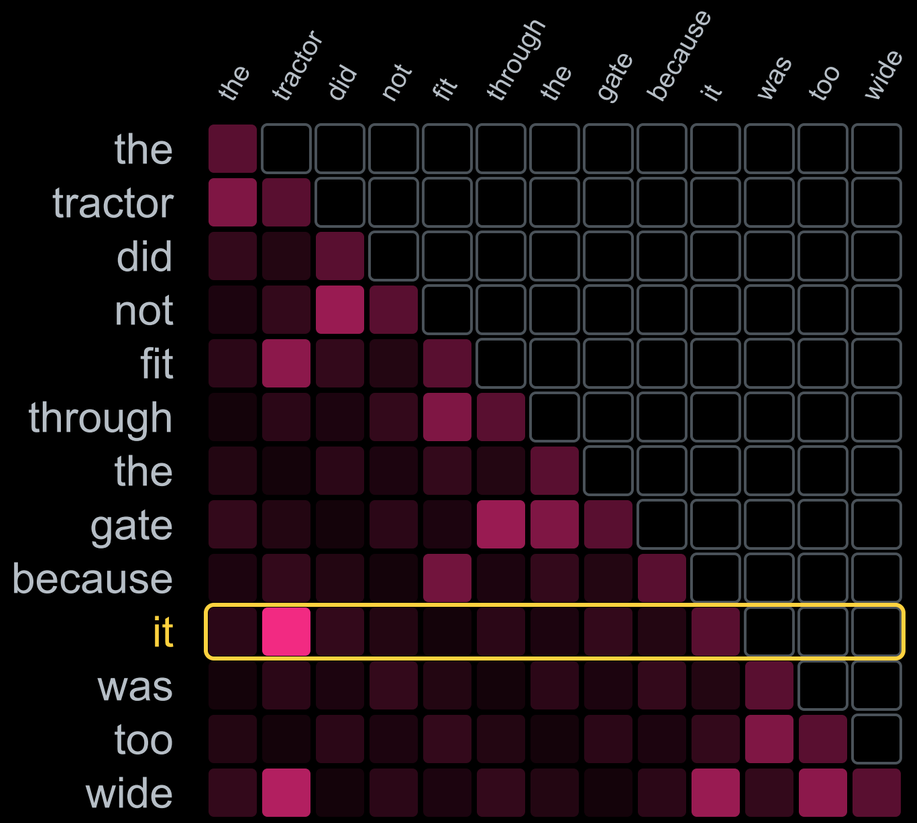
13

cells

169

words instead of tokens and strengths for illustration

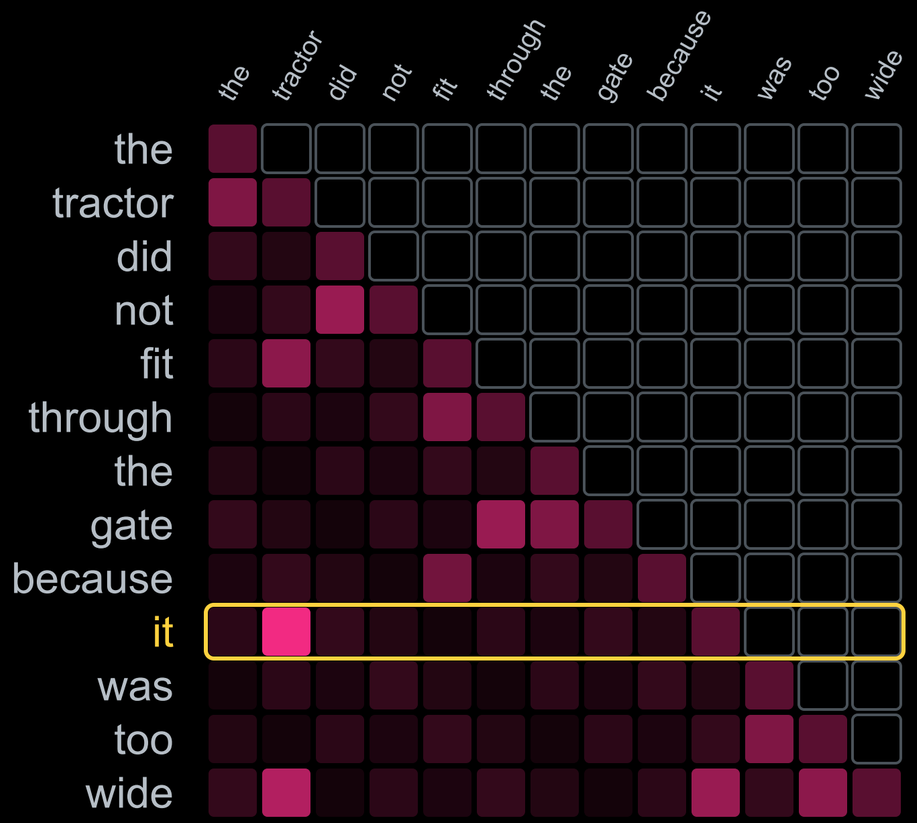
every token with every token



tokens	cells
13	169
1,000	1,000,000

words instead of tokens and strengths for illustration

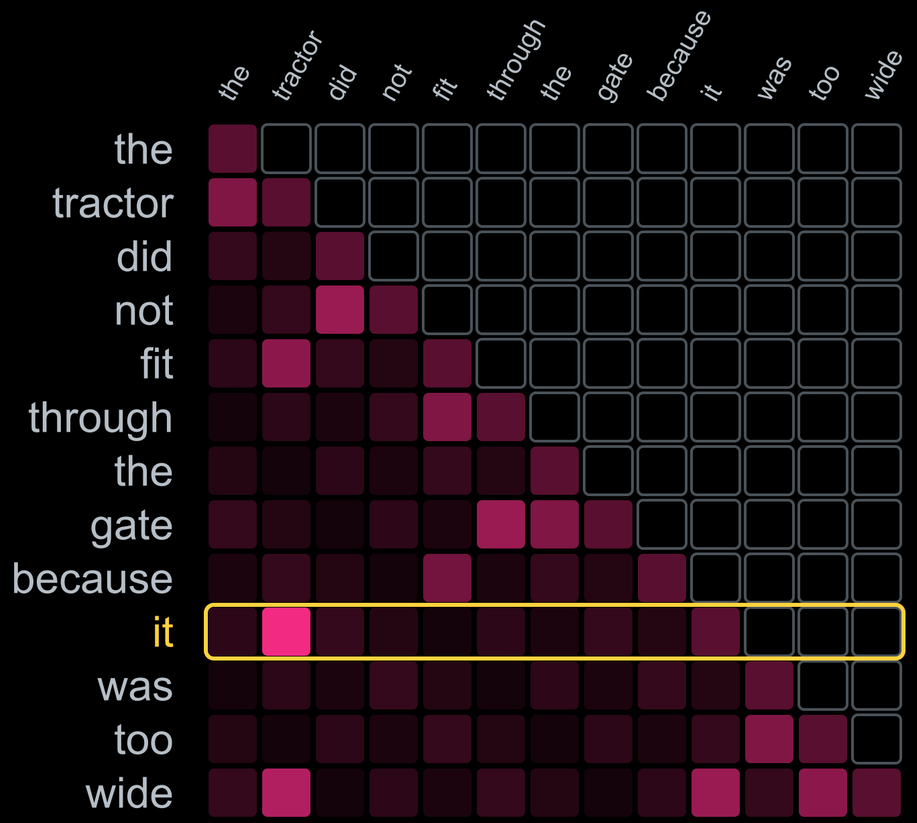
every token with every token



tokens	cells
13	169
1,000	1,000,000
131,072	17,179,869,184

words instead of tokens and strengths for illustration

every token with every token



tokens	cells
13	169
1,000	1,000,000
131,072	17,179,869,184

twice the context means four times the cells.

words instead of tokens and strengths for illustration

a language model writes one token at a time, and the whole context flows into the prediction of every single one.



image: ai-generated (gpt-image-2.5-flare)