Review of language models

CS 221M
Week 1, Lecture 2

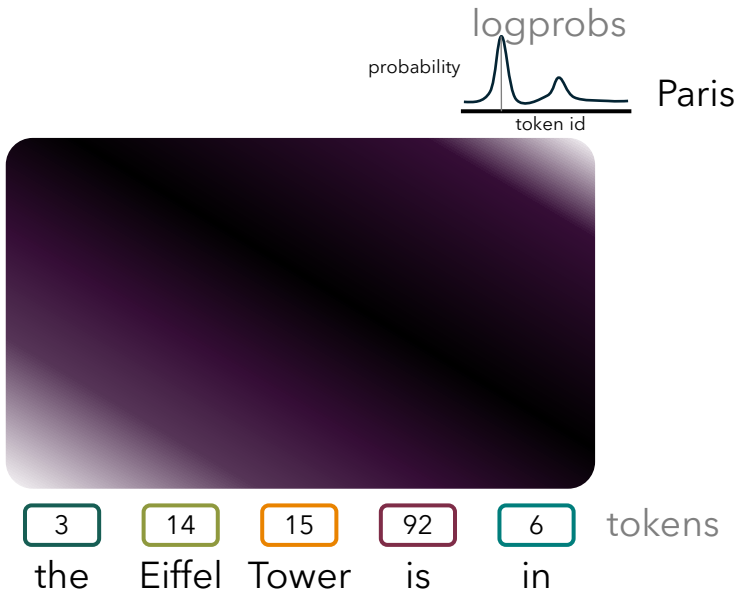
cs221m.stanford.edu

Date	Lesson	Readings	Materials
Week 1 Mon. March 30	Introduction	—	—
Week 1 Wed. April 1	Review of language models	Rush et al. 2018 annotated transformer Elhage et al. 2021 circuits	—
Week 2 Mon. April 6	Behavioral analysis and input attribution	Jia and Liang 2017 adversarial evaluation Sundararajan et al. 2017 integrated gradients	—
Week 2 Wed. April 8	Probes for decoding activations	Wendler et al. 2024 llamas think in english Tenney et al. 2019 BERT rediscovers NLP pipeline Marks et al. 2023 geometry of truth	—

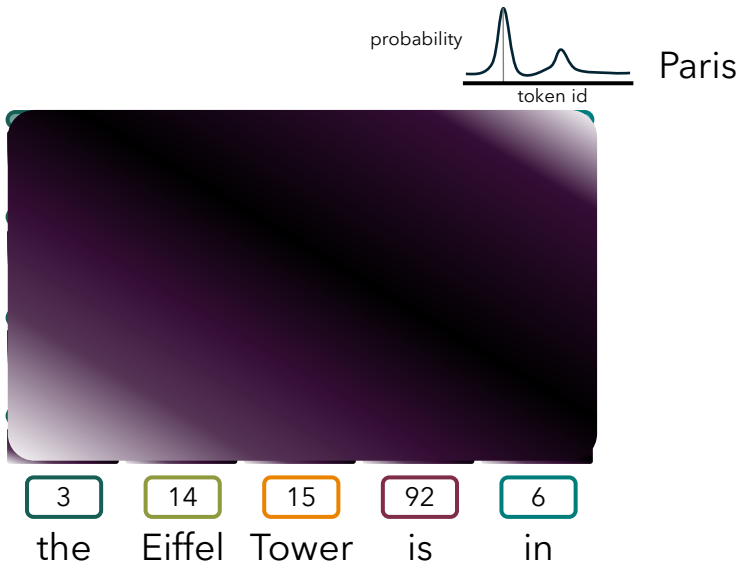
Paris

the Eiffel Tower is in

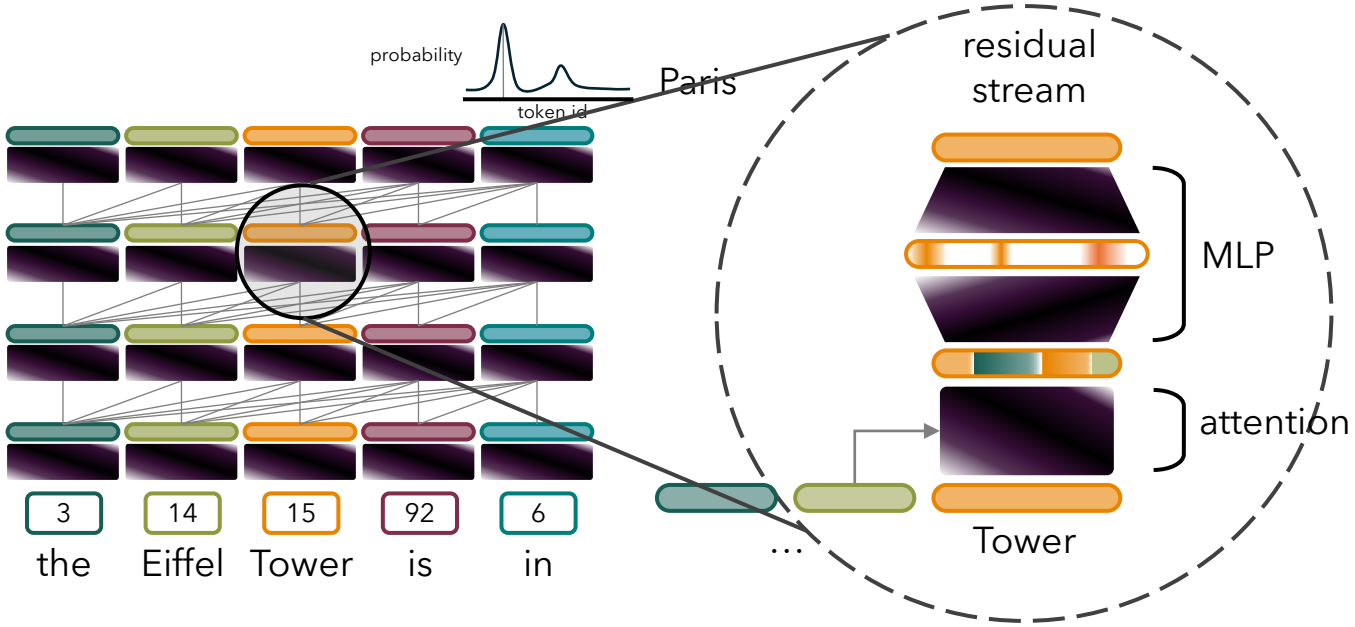
The gameplan



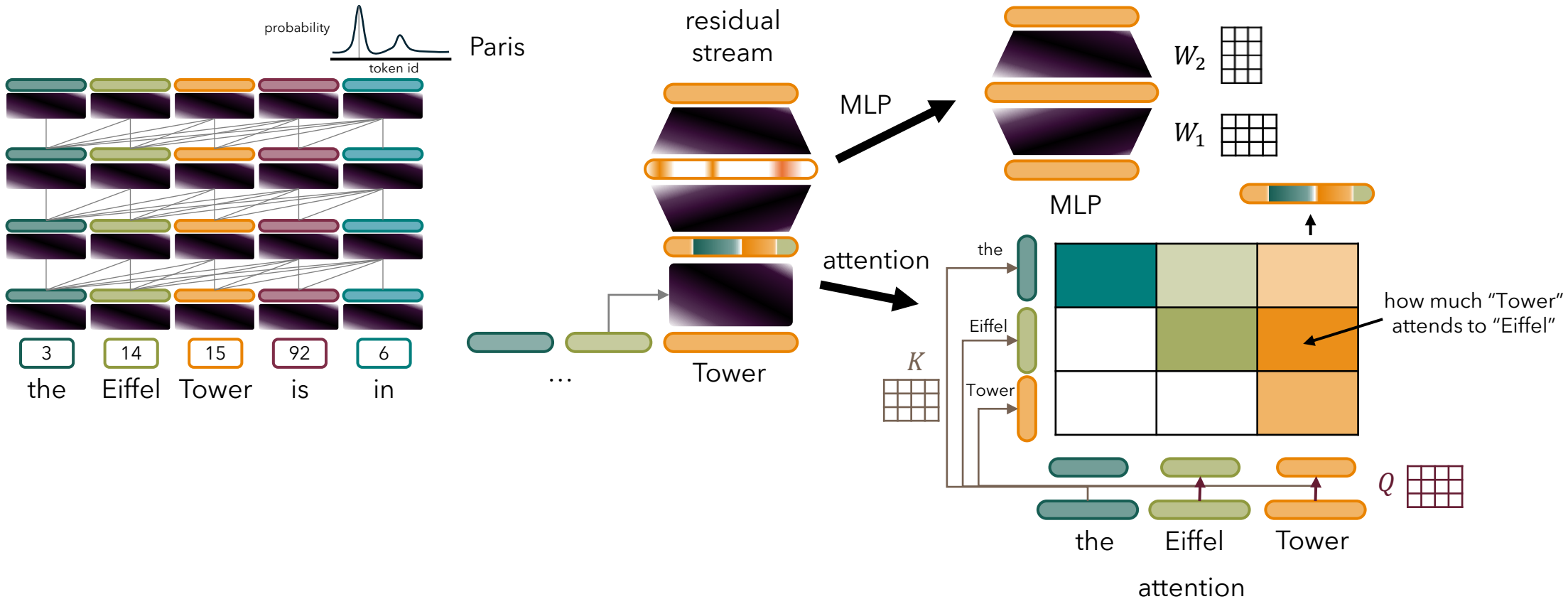
The gameplan



The gameplan



The gameplan



The gameplan

What's a language model?

- Autoregressive
  - Generates one token at a time

What's a language model?

- Autoregressive
  - Generates one token at a time

What's a language model?

- Autoregressive
  - Generates one token at a time

What's a language model?

- Autoregressive
  - Generates one token at a time

What's a language model?

- Autoregressive
  - Generates one token at a time

What's a language model?

- Autoregressive
  - Generates one token at a time

What's a language model?

- Autoregressive
  - Generates one token at a time

What's a language model?

- Autoregressive
  - Generates one token at a time

What's a language model?

- Autoregressive
  - Generates one token at a time

What's a language model?

- A

Paris

Don't think, just give me the answer.

-20, -19, 88, 29, 84, 80, 18, 92, 87, 52, 16, 60, 76, -75, 30, 84, -82, -4, 63, 38, -60, -54, -91, -98, 87, 75, 64, -85, 48, -93, -57, -14, -9, -88, -12, 1, -84, -73, 88, 86, 63, 68, 54, 42, 21, 9, -31, -92, -34, 1

What's the largest negative number?

❌ qwen3-vl-235b-a22b-instruct

-1

https://arena.ai

The Eiffel Tower is in

What's a language model?

- A

Paris

Don't think, just give me the answer. Find the largest negative number in the list.

-20, -19, 88, 29, 84, 80, 18, 92, 87, 52, 16, 60, 76, -75, 30, 84, -82, -4, 63, 38, -60, -54, -91, -98, 87, 75, 64, -85, 48, -93, -57, -14, -9, -88, -12, 1, -84, -73, 88, 86, 63, 68, 54, 42, 21, 9, -31, -92, -34, 1

What's the largest negative number?

qwen3-vl-235b-a22b-instruct

-4

https://arena.ai

The Eiffel Tower is in

What's a language model?

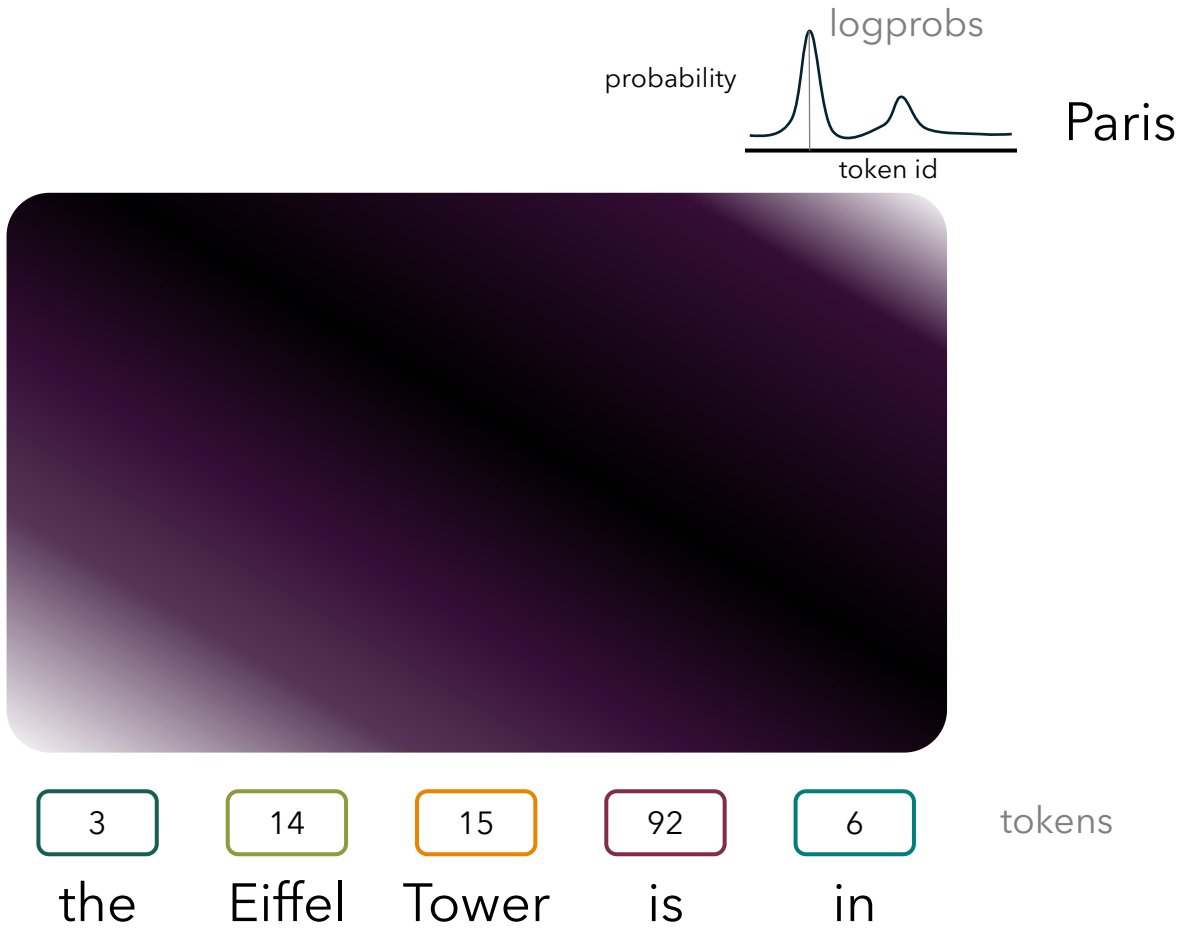
- Autoregressive
  - Generates one token at a time
  - Processes one token at a time
- Pre-trained
  - On next-token completion across internet

One layer back: tokens and probabilities

Paris

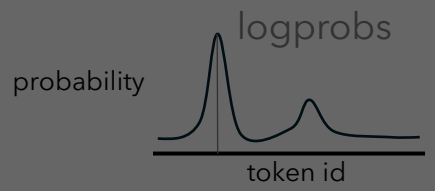
the Eiffel Tower is in

One layer back: tokens and probabilities

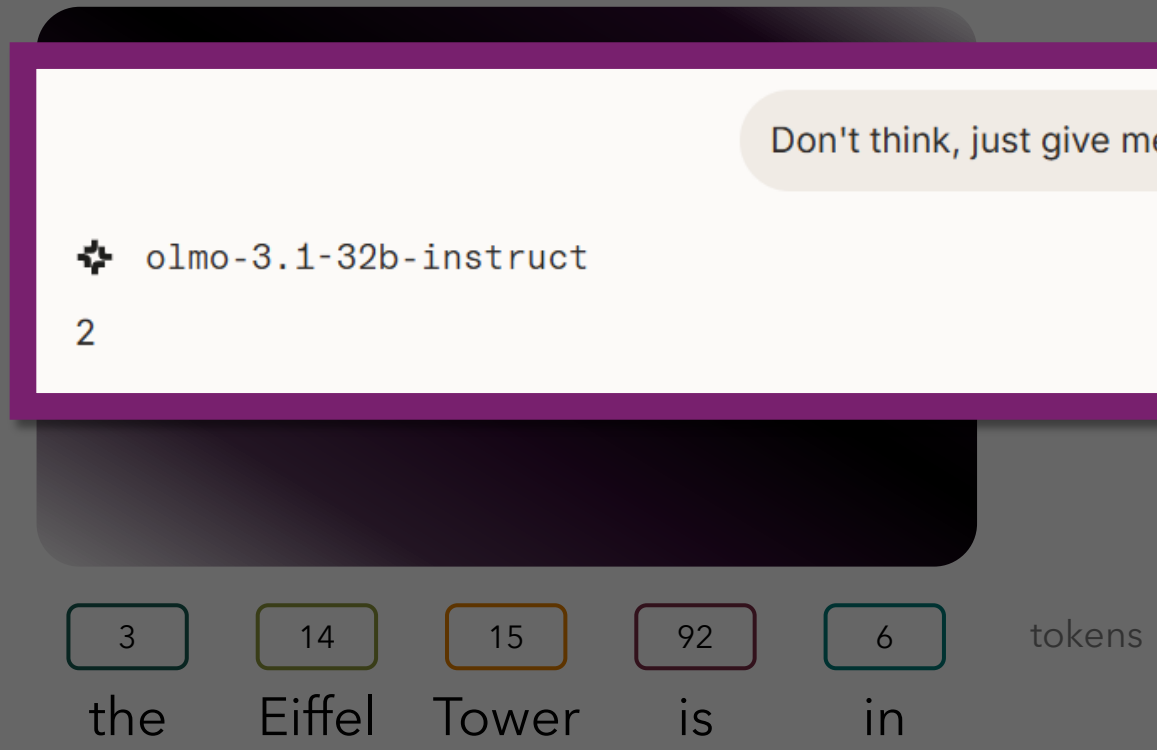


- Input
  - Text gets tokenized into words or subwords
  - Each token has a unique number ID
- Output
  - Probability distribution over the next token
  - Sample one token from this distribution

One layer back: tokens and probabilities



Paris



- Input
  - Text gets tokenized

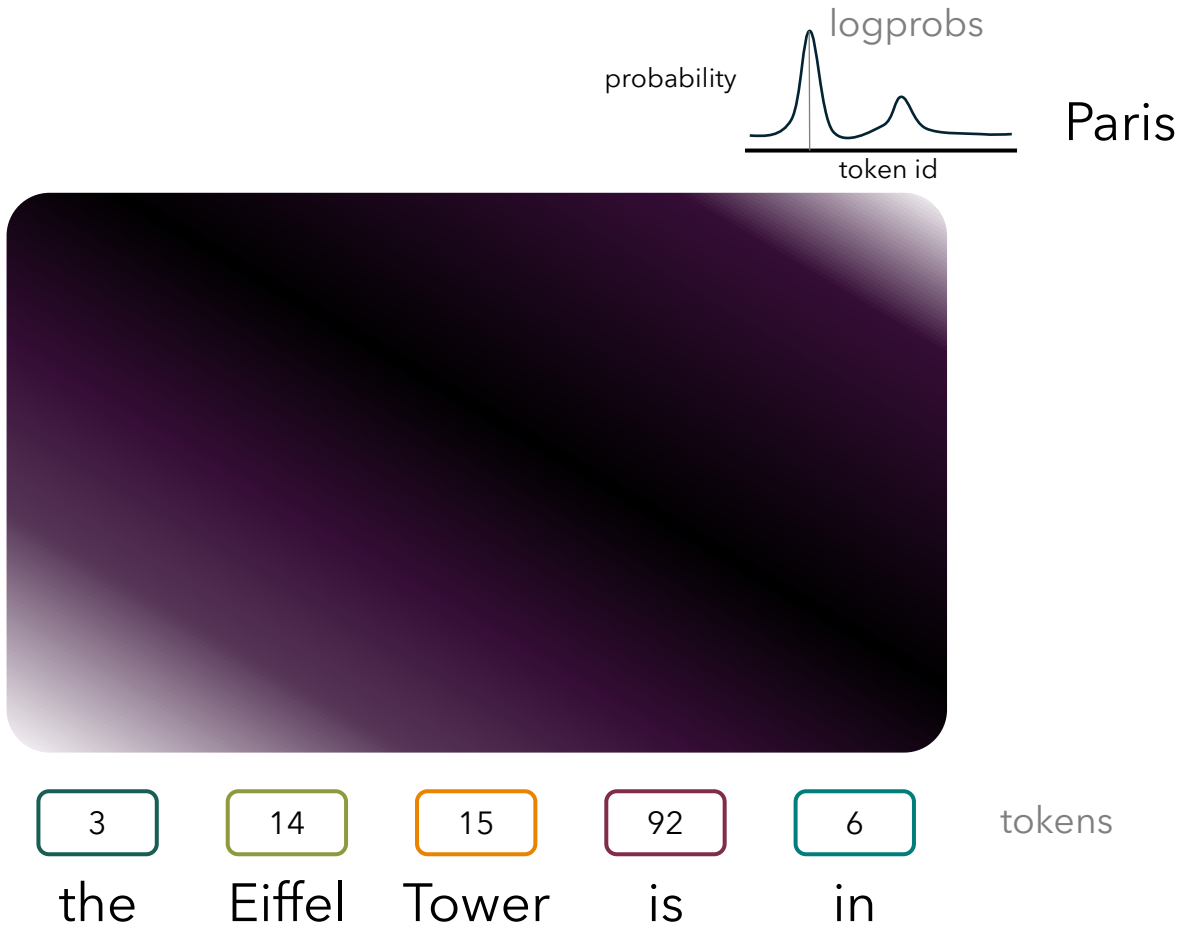
Don't think, just give me the final answer. How many Rs are in STRAWBERRY?

olmo-3.1-32b-instruct

2

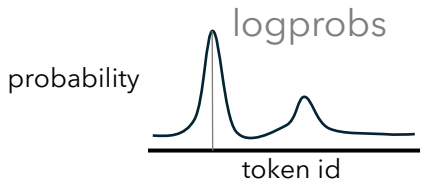
- Output
  - Probability distribution over the next token
  - Sample one token from this distribution

One layer back: tokens and probabilities

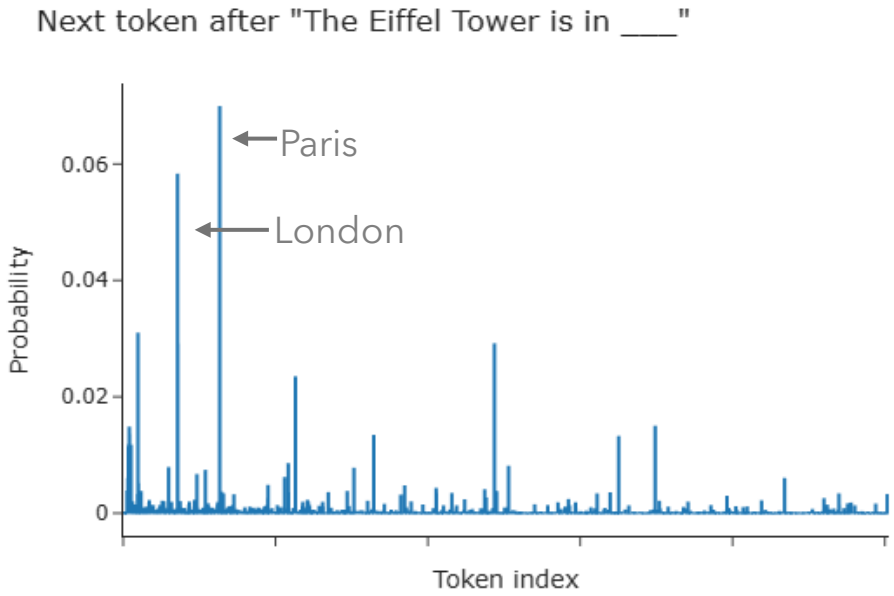


- Input
  - Text gets tokenized into words or subwords
  - Each token has a unique number ID
- Output
  - Probability distribution over the next token
  - Sample one token from this distribution

One layer back: tokens and probabilities

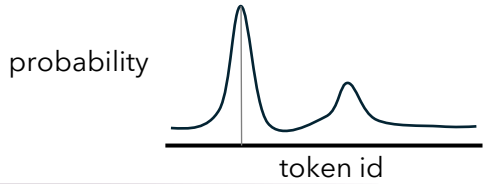


Paris



3	14	15	92	6	tokens
the	Eiffel	Tower	is	in	

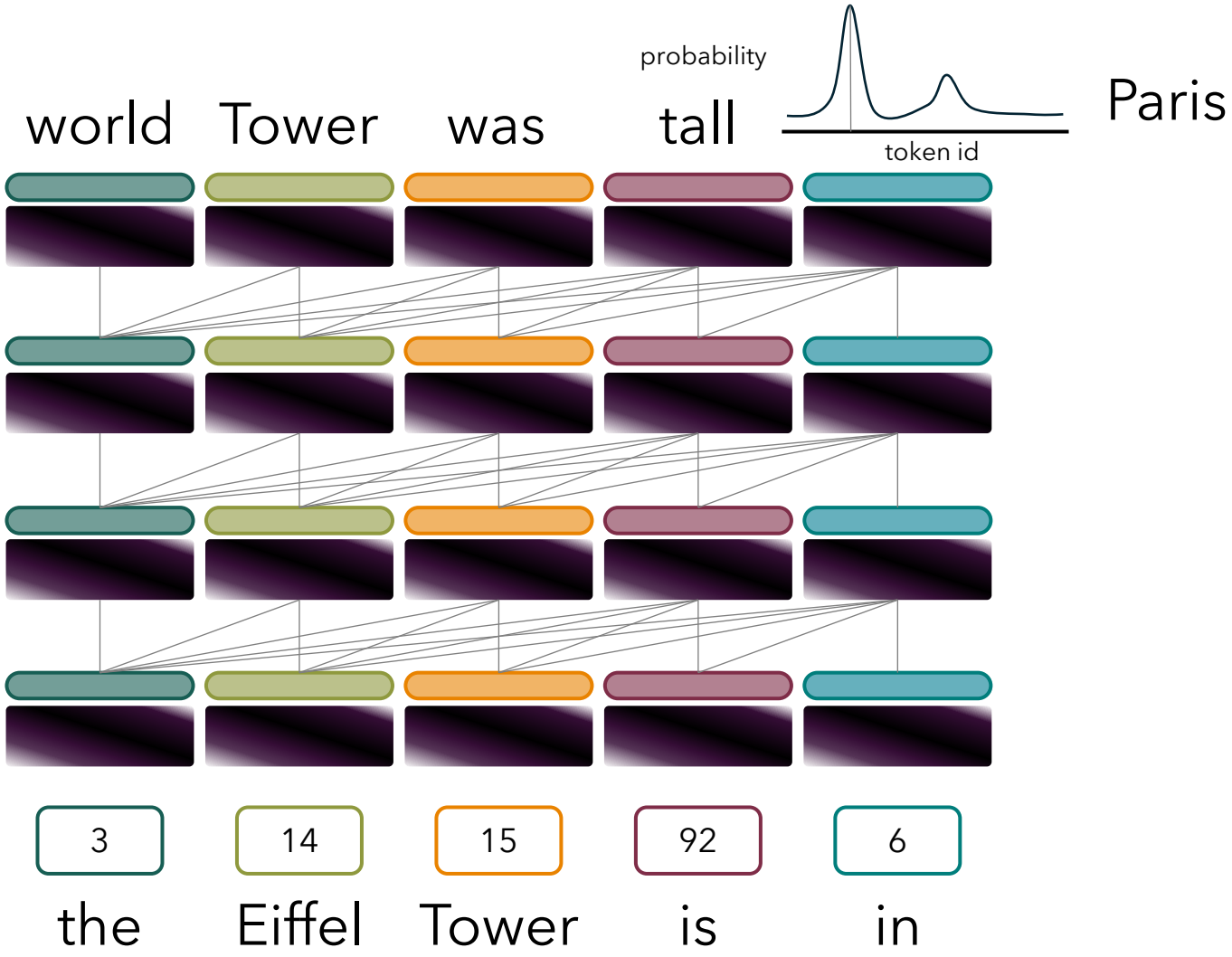
Inside the language model



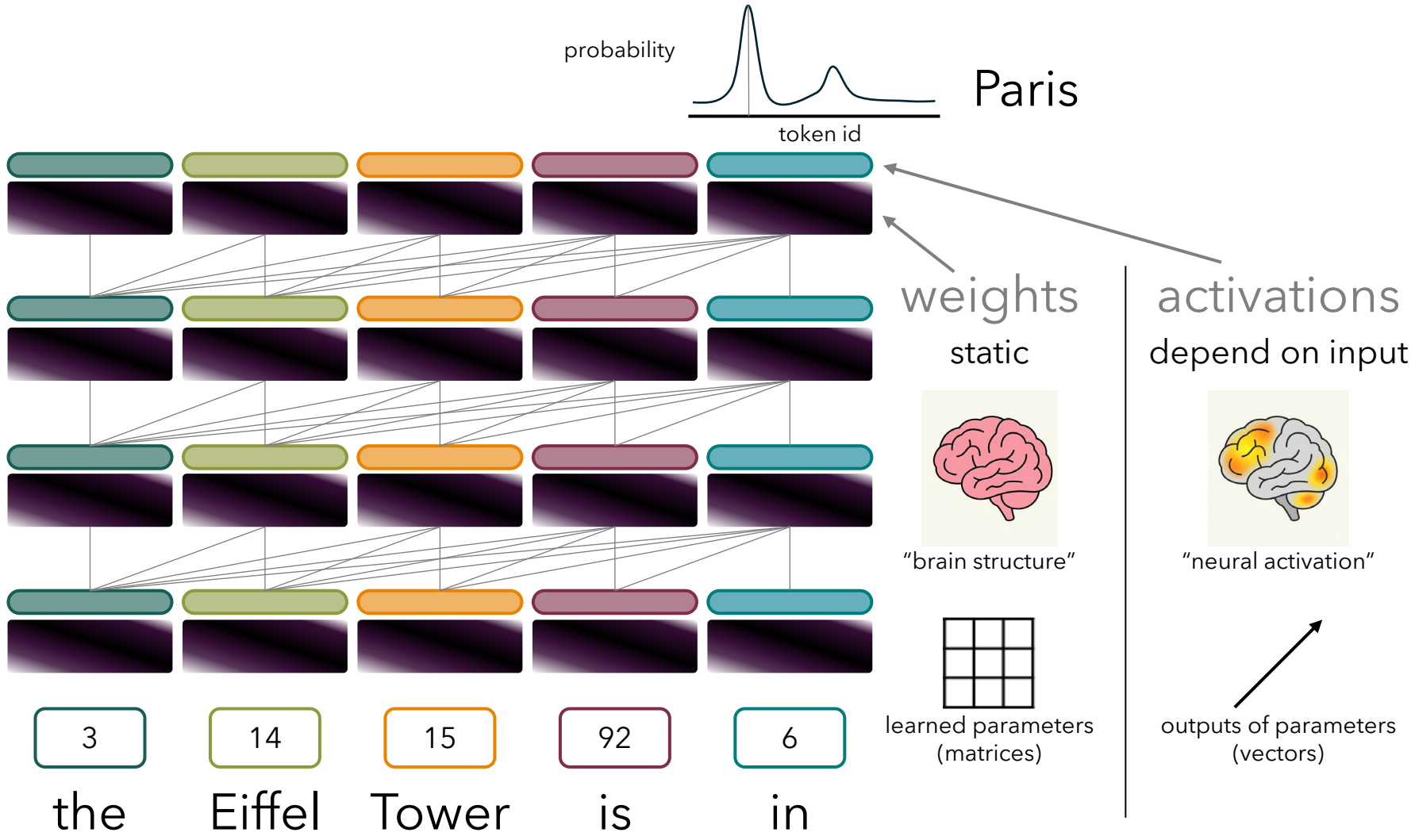
Paris

3	14	15	92	6
the	Eiffel	Tower	is	in

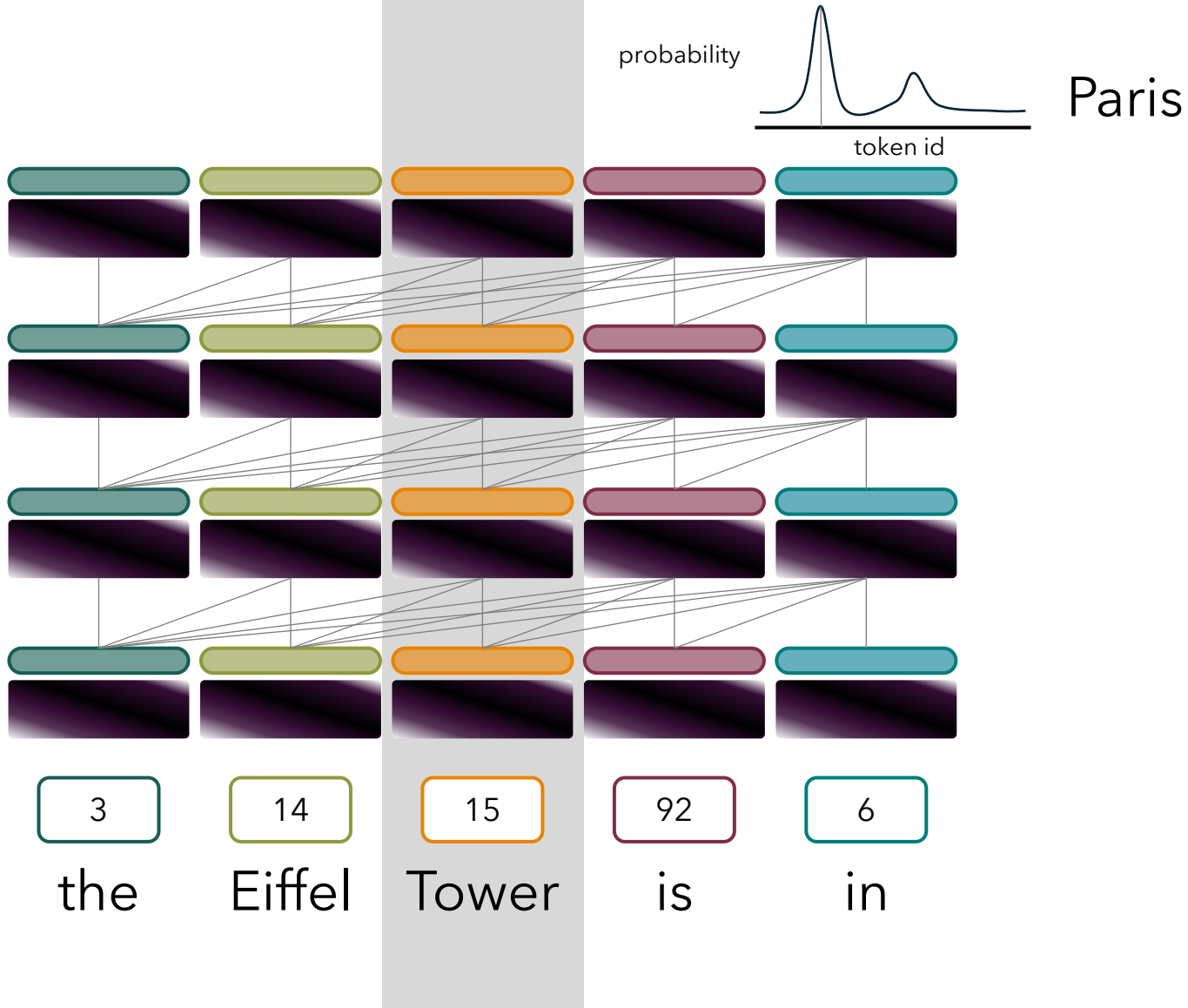
Parallel processing



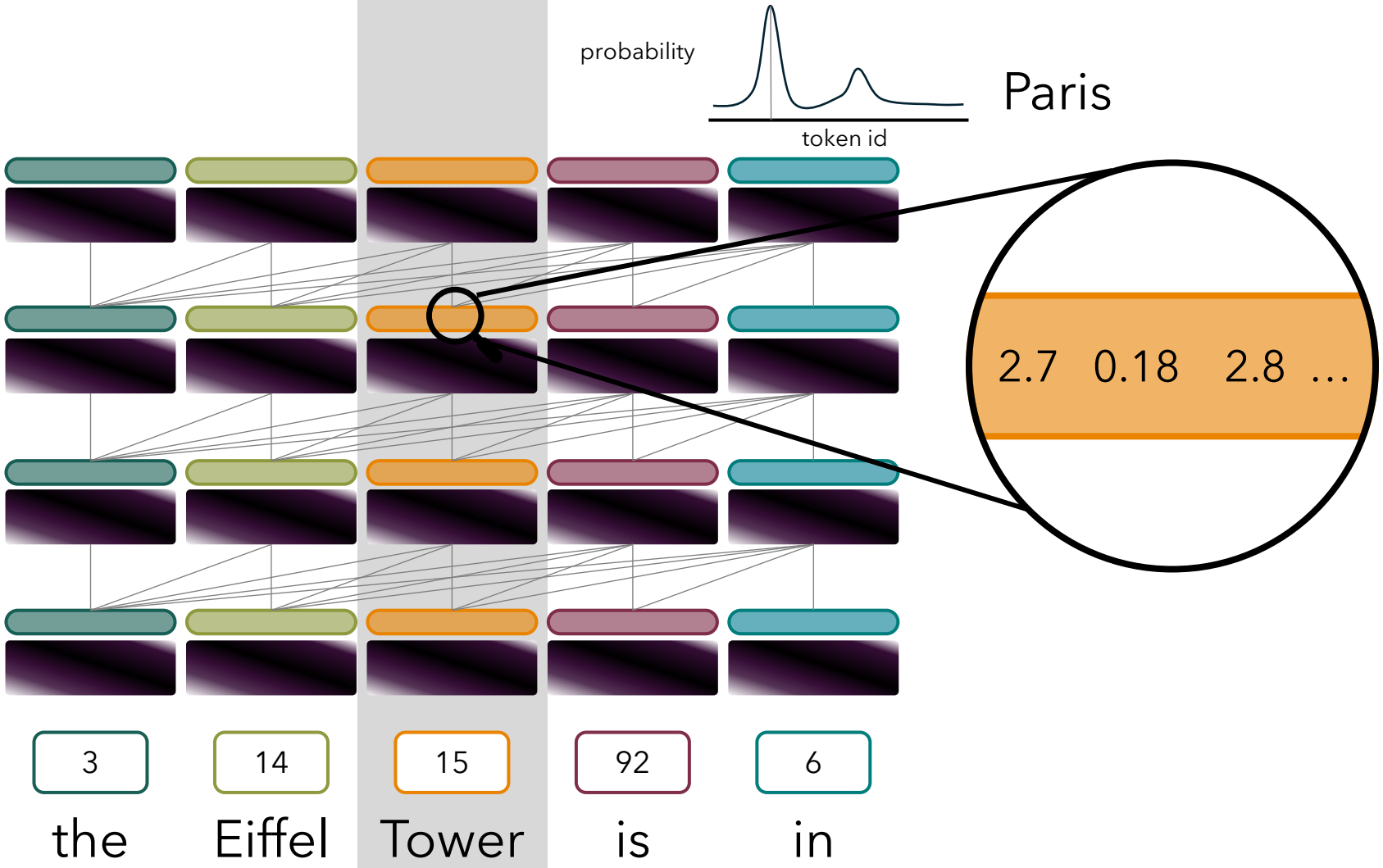
Weights vs. activations



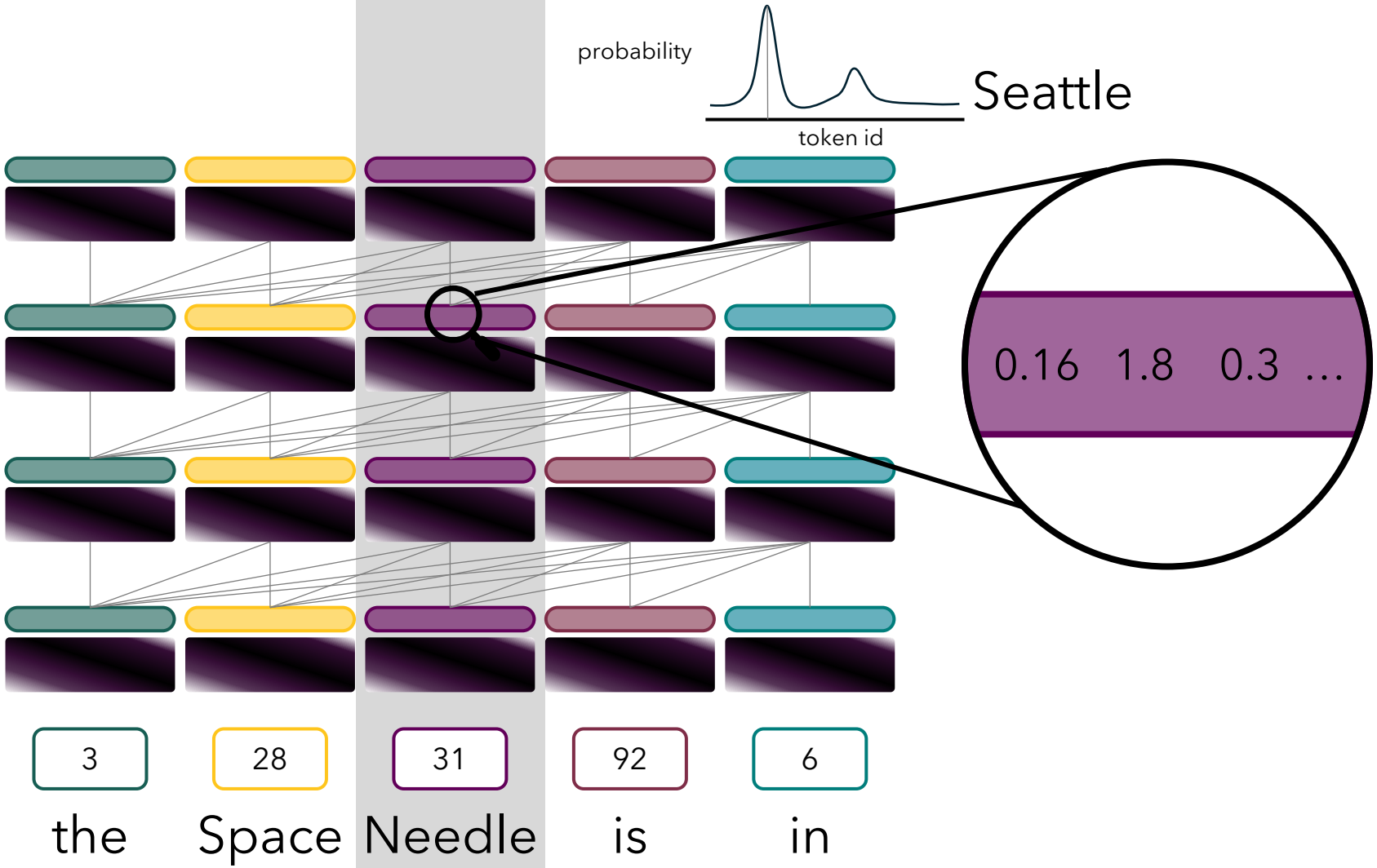
The residual stream



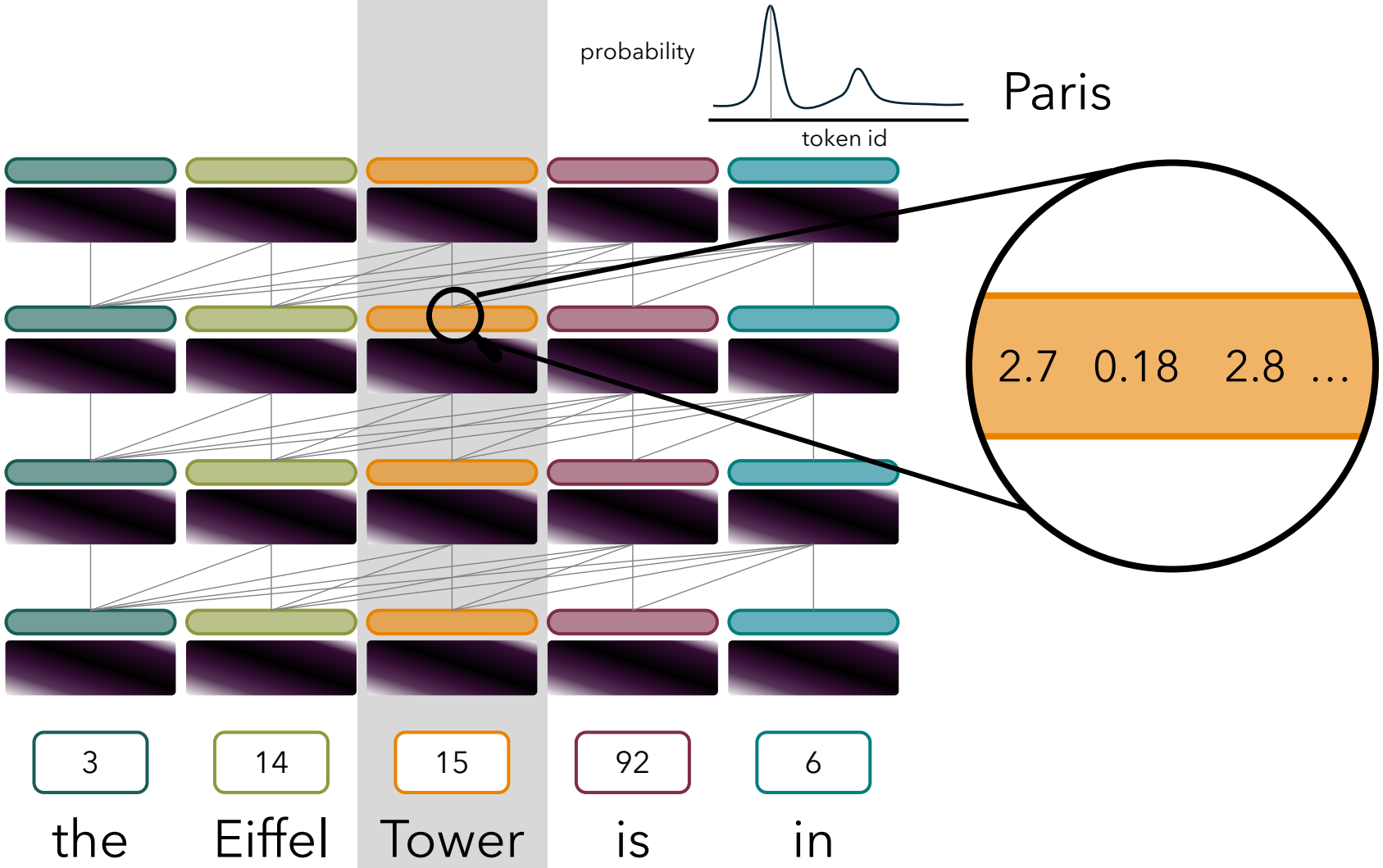
The residual stream



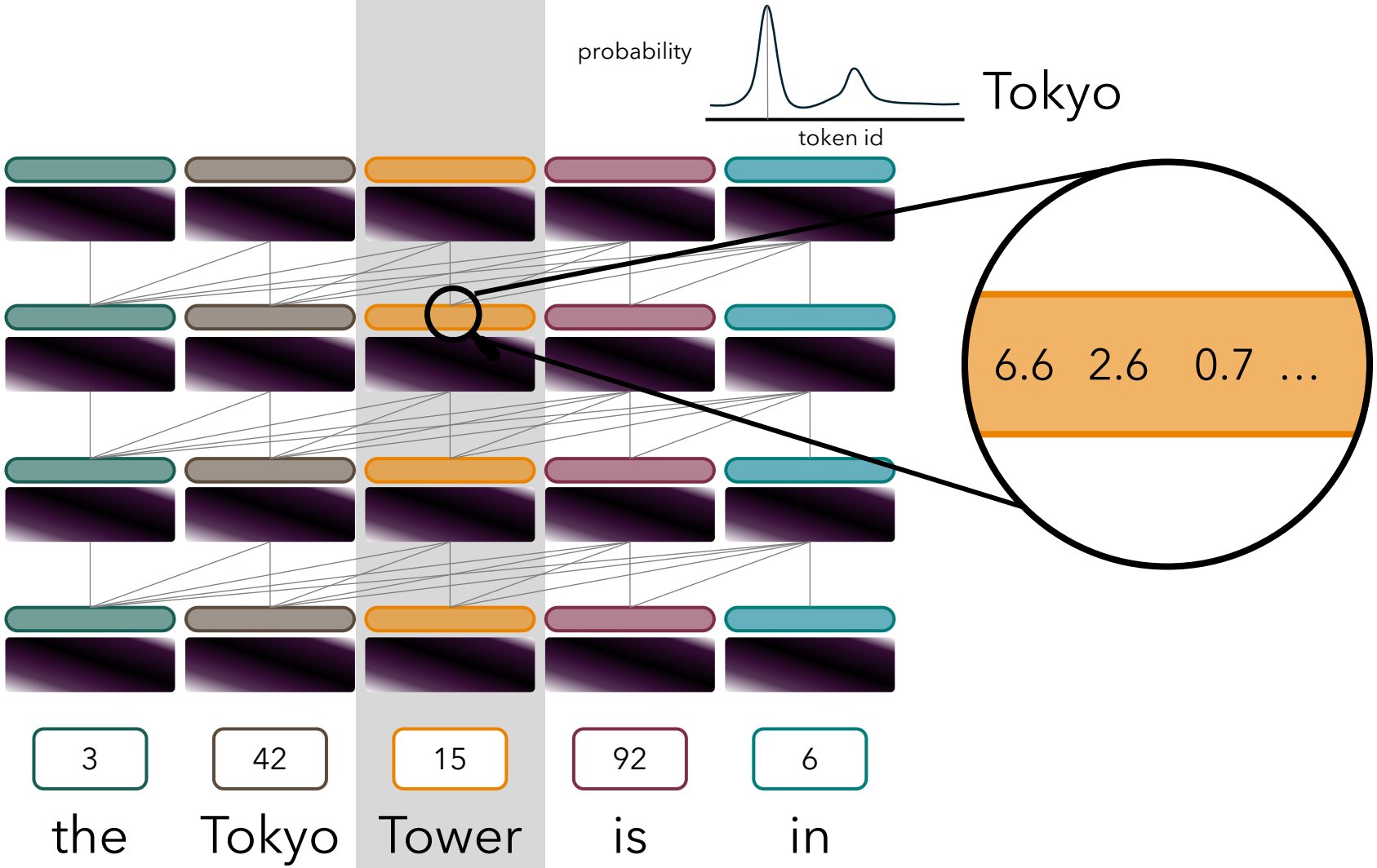
The residual stream



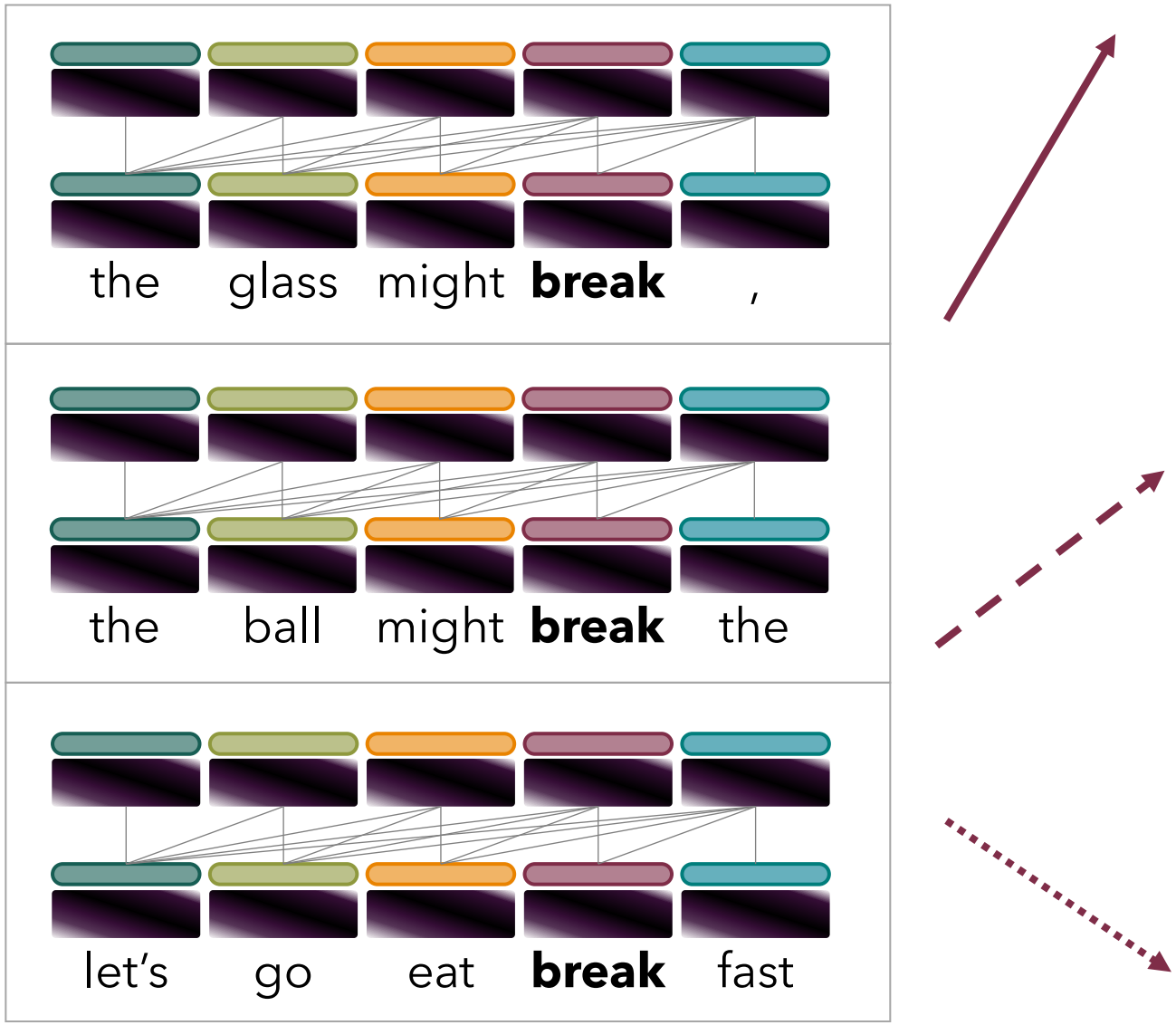
The residual stream



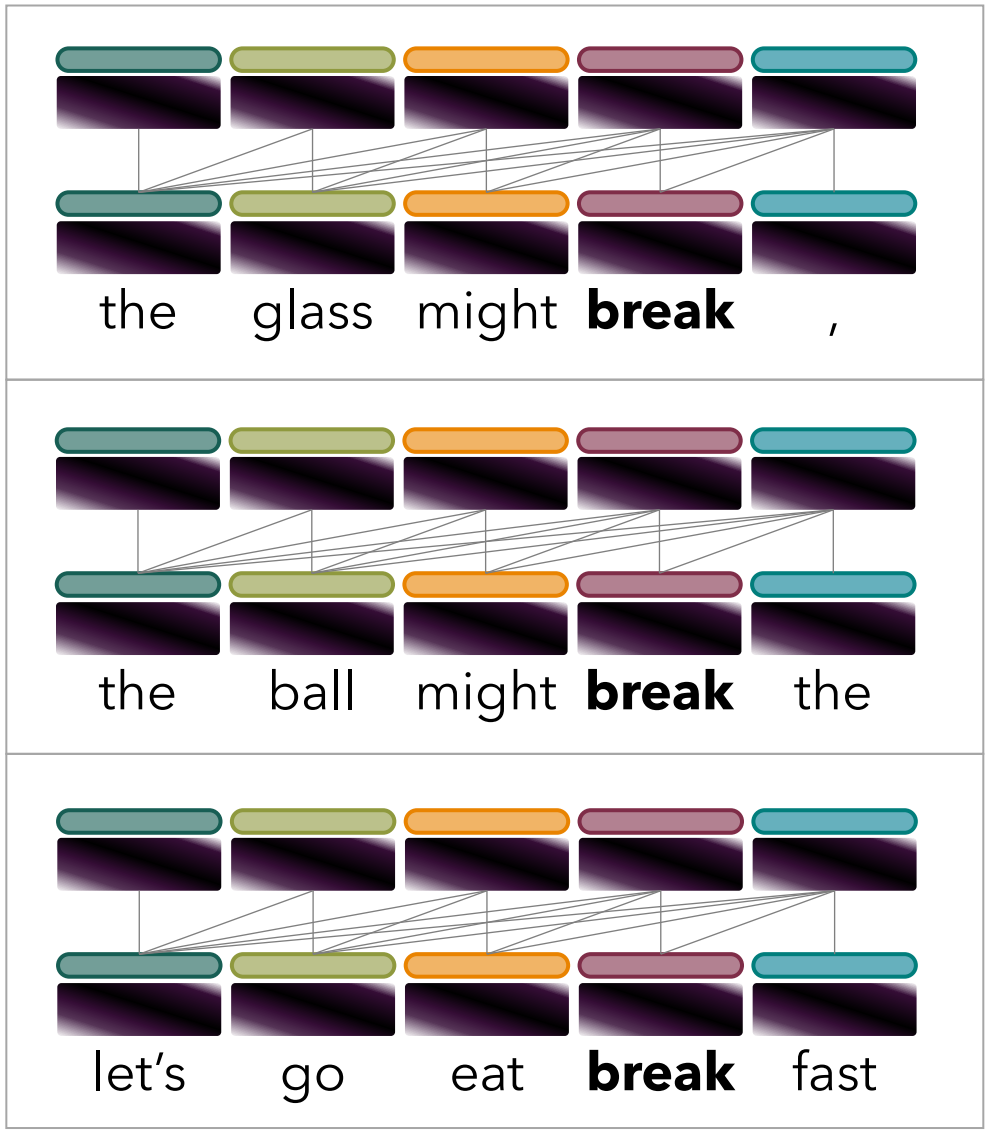
The residual stream

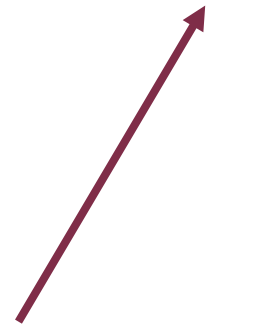


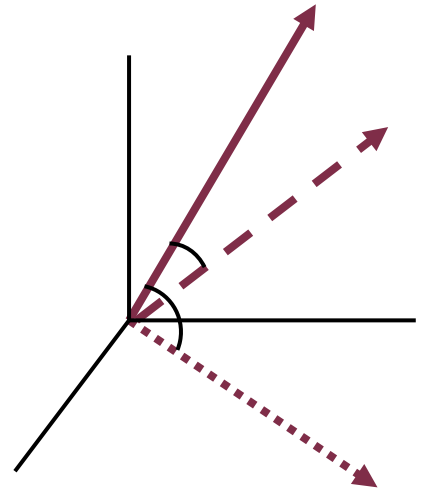
The residual stream



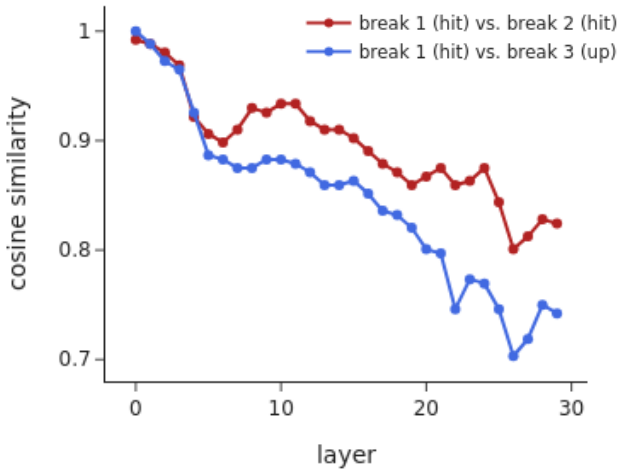
The residual stream





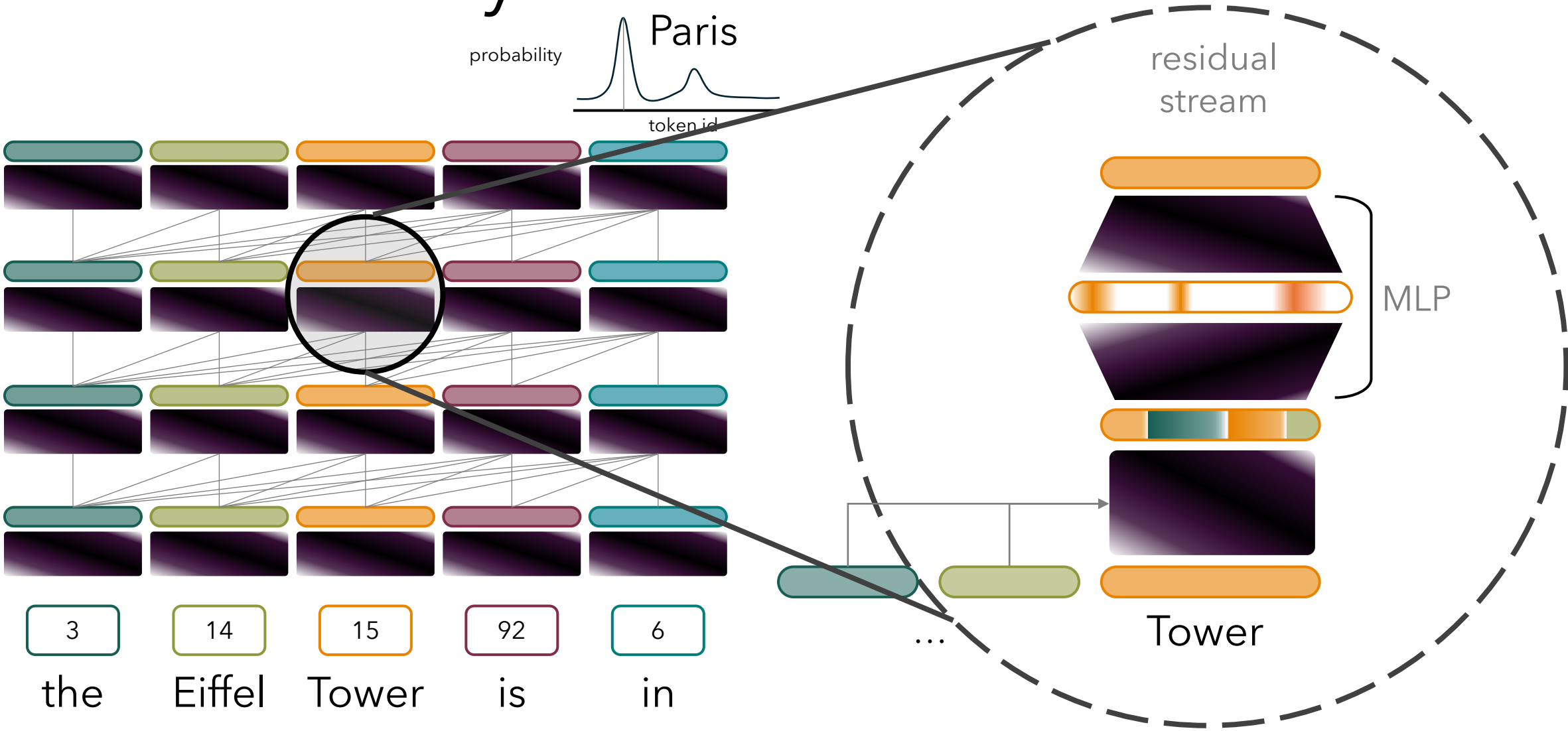


similarity between activations ~
similarity between tokens in that context

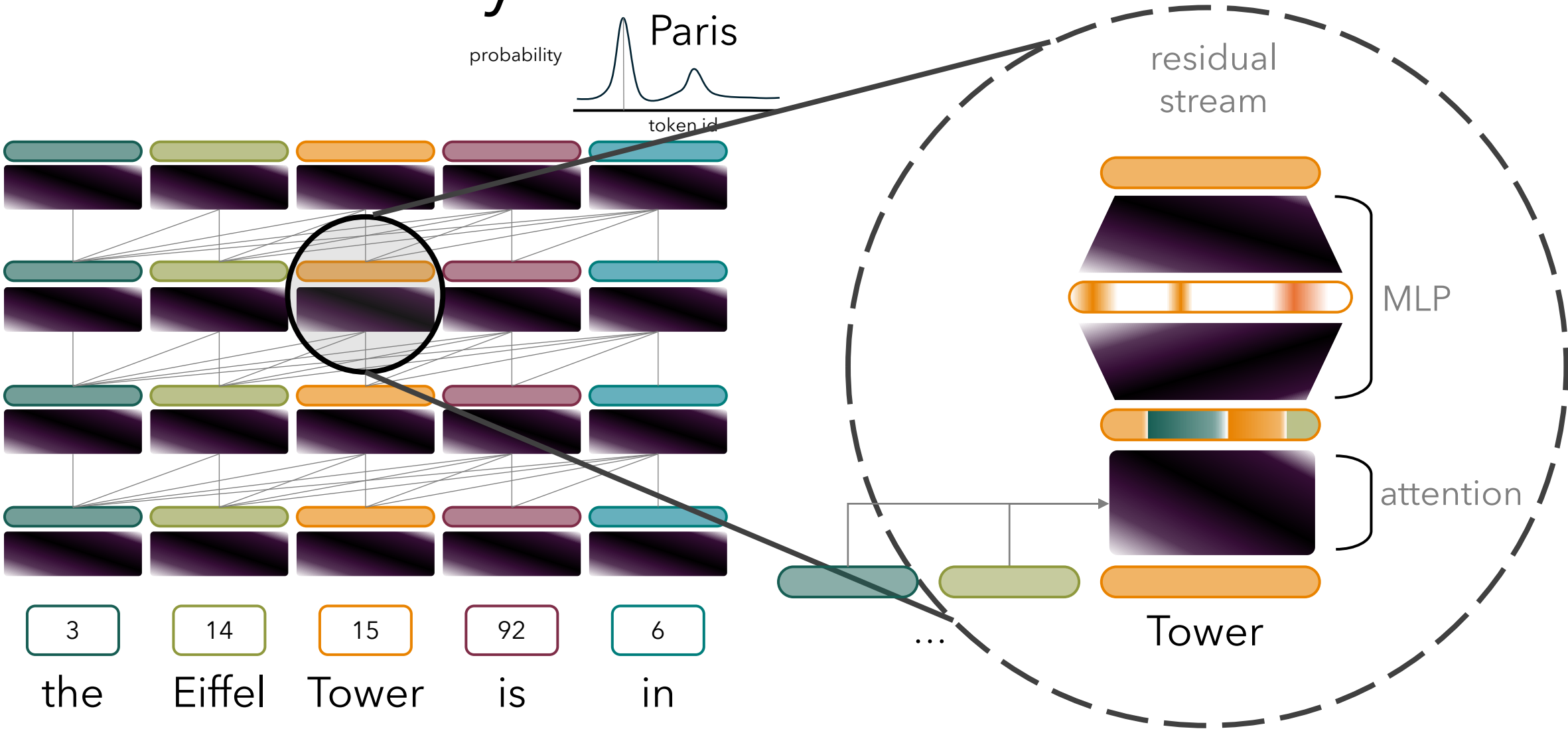


earlier layers: syntactic
later layers: semantic
last layer: next token prediction

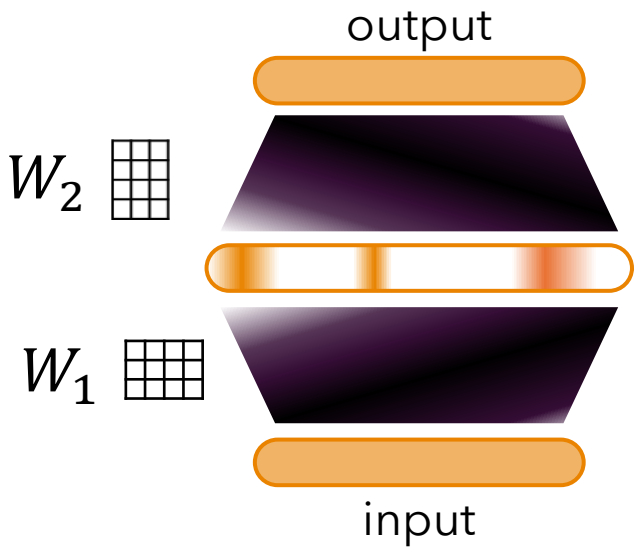
Inside each layer



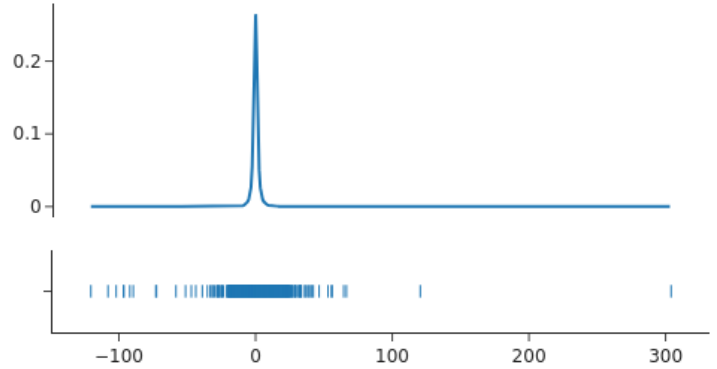
Inside each layer

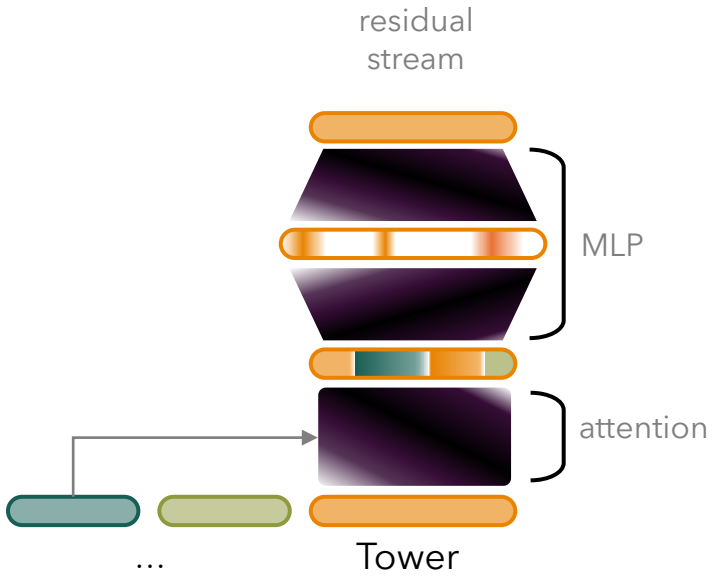


Inside each layer: MLP



distribution of MLP activations



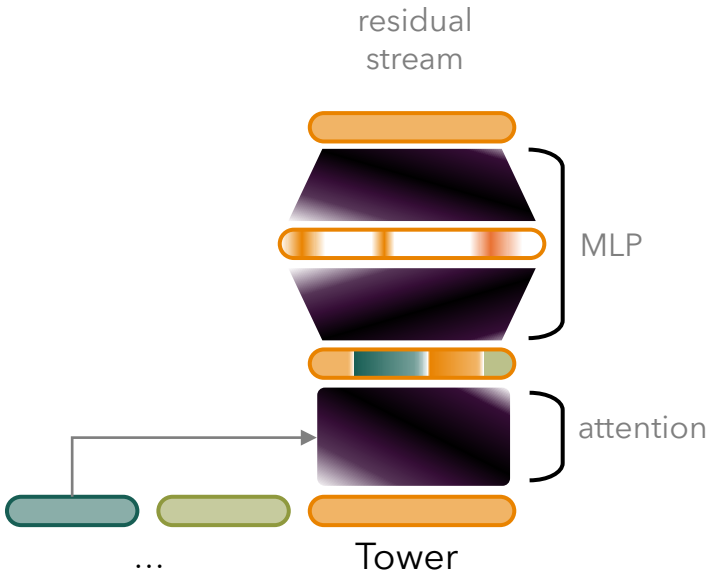


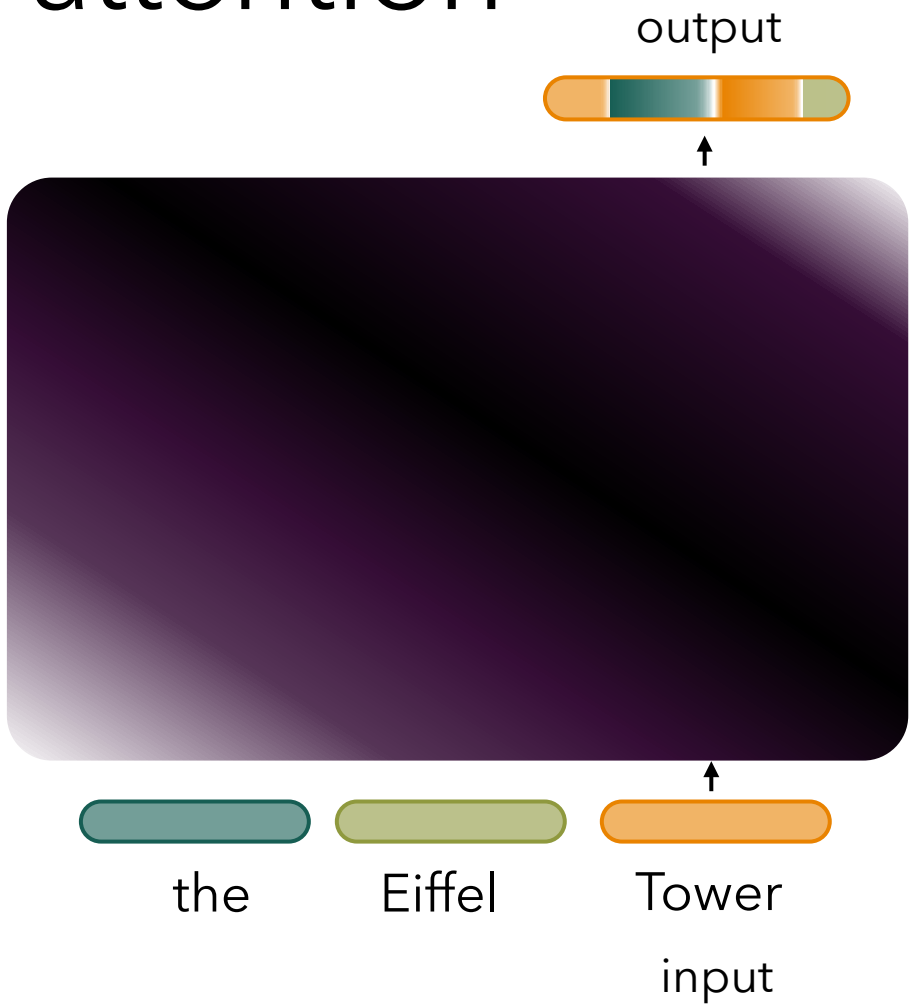
local	dense	sparse
• only information from current token	• makes up ~70% of the model	• few neurons with very large values

Inside each layer: attention

moves information

- only way to move information across tokens!

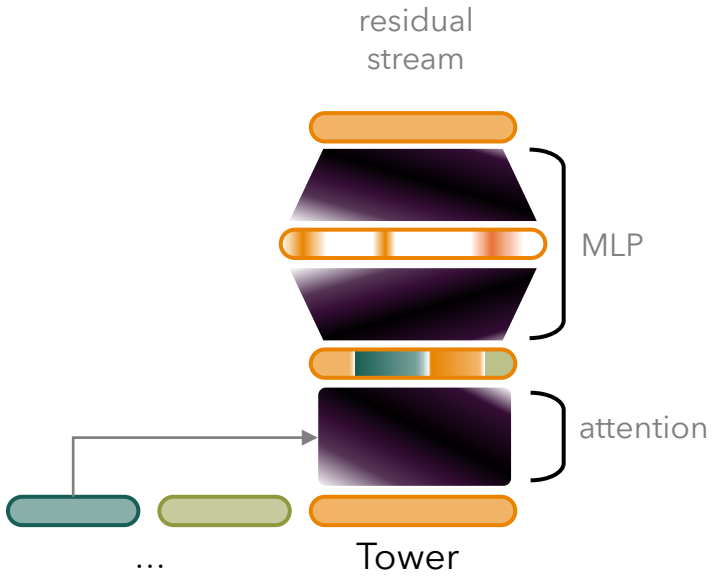


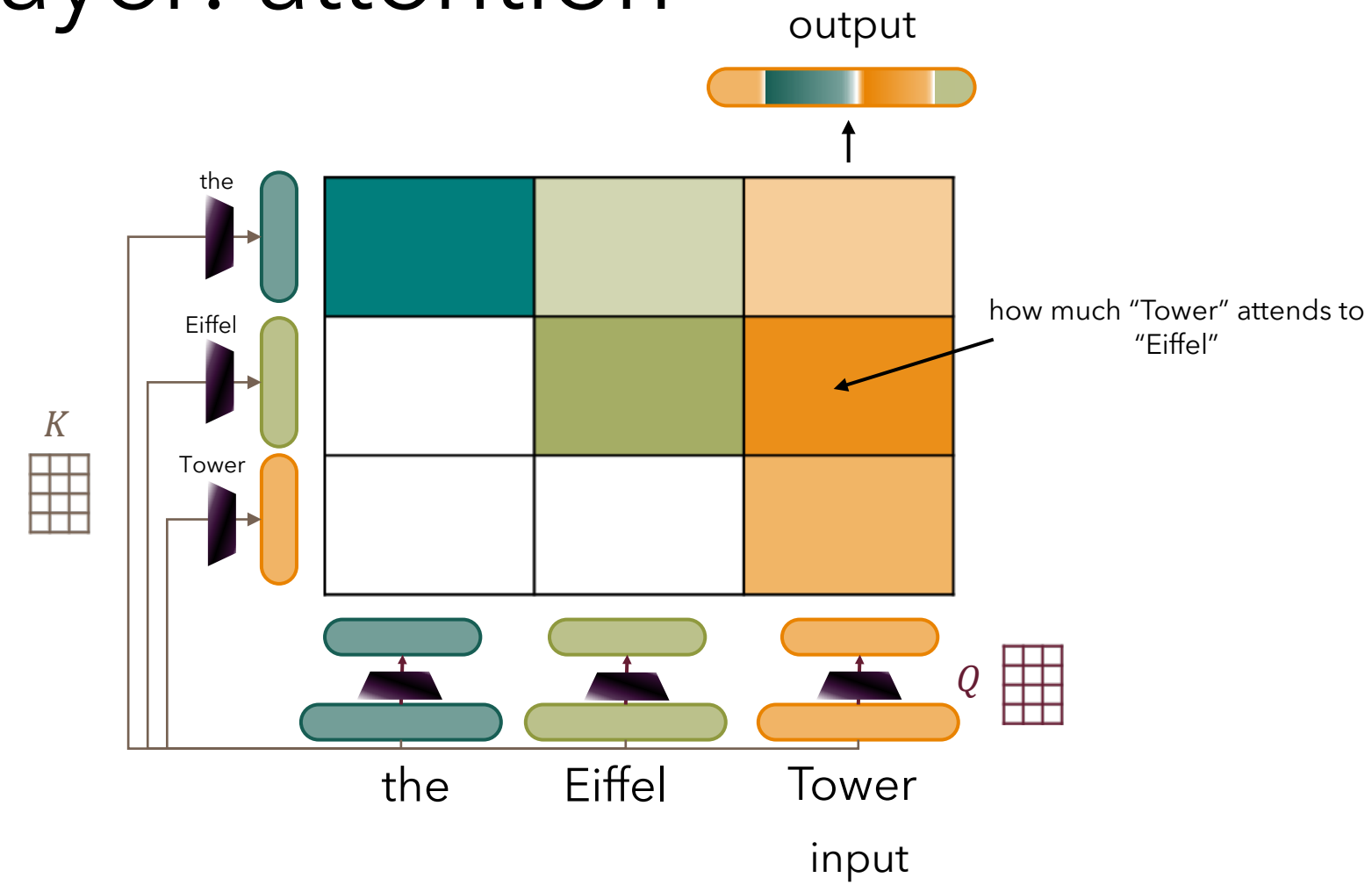


Inside each layer: attention

moves information

- only way to move information across tokens!

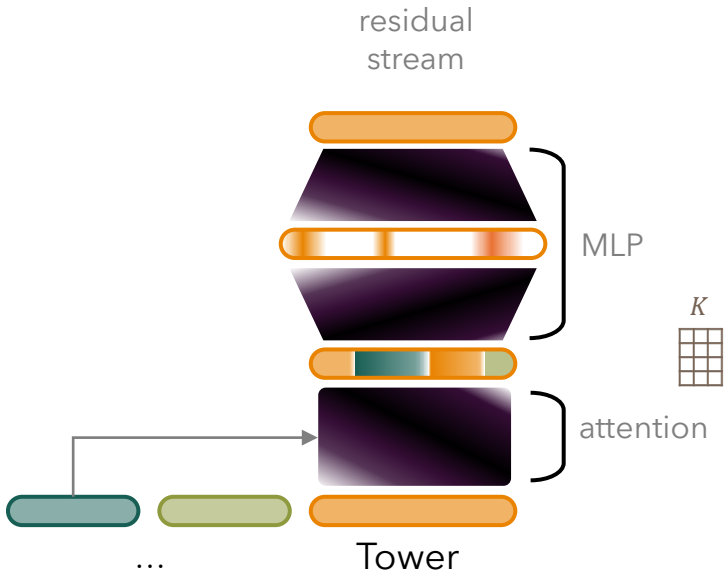


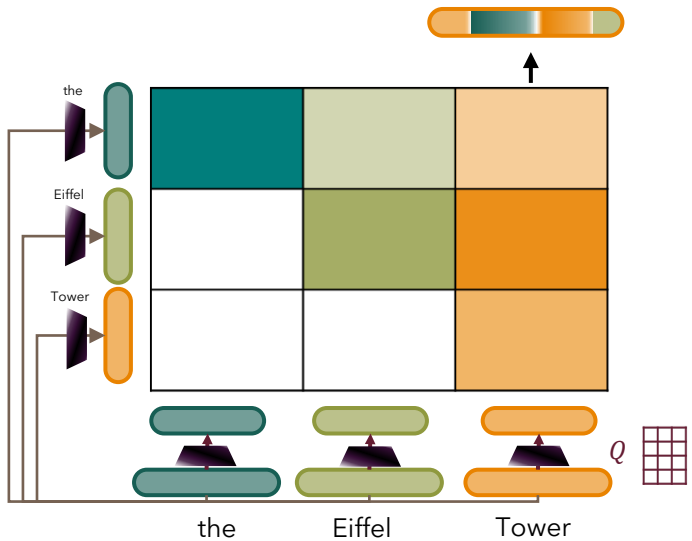


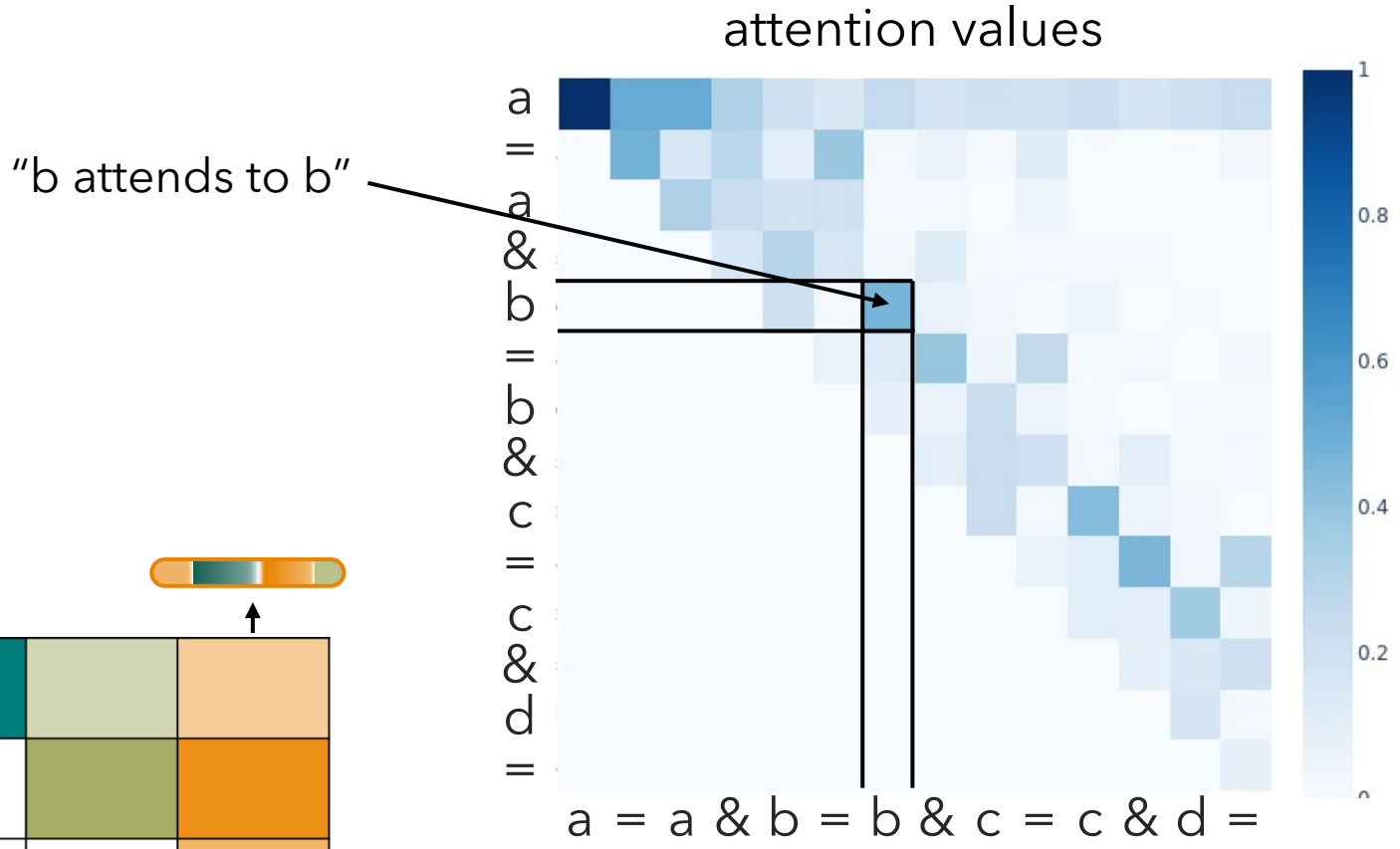
Inside each layer: attention

moves information

- only way to move information across tokens!

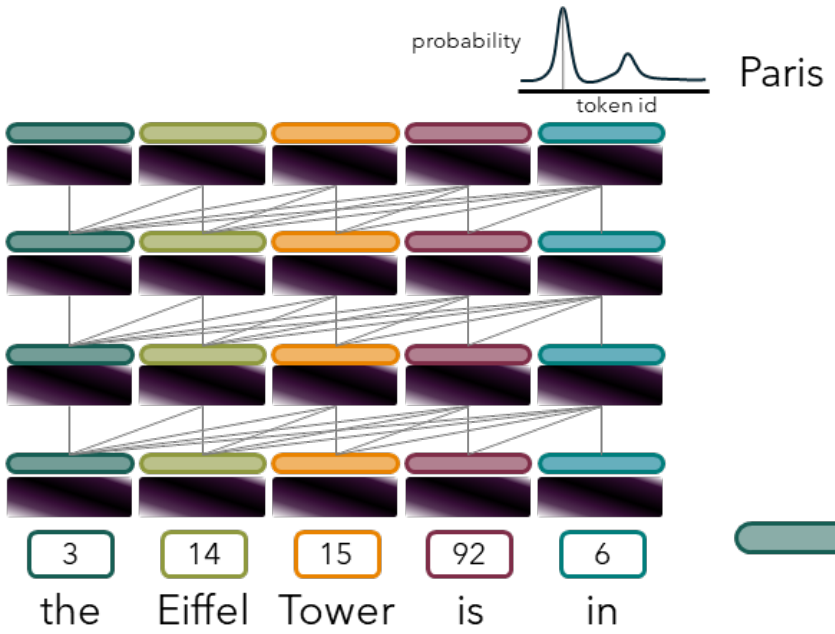


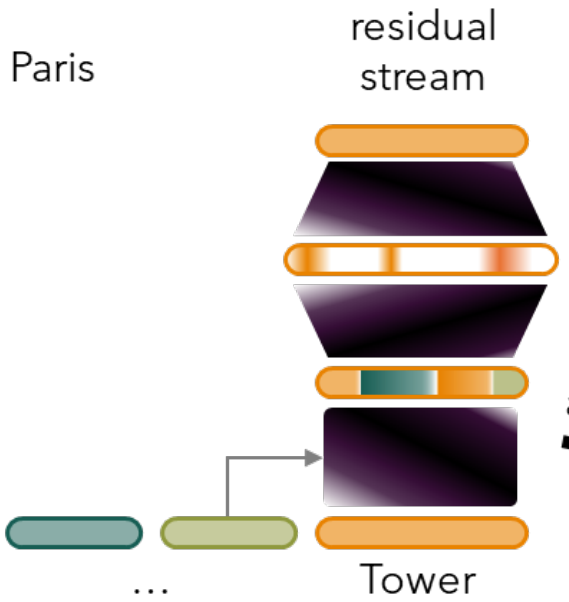


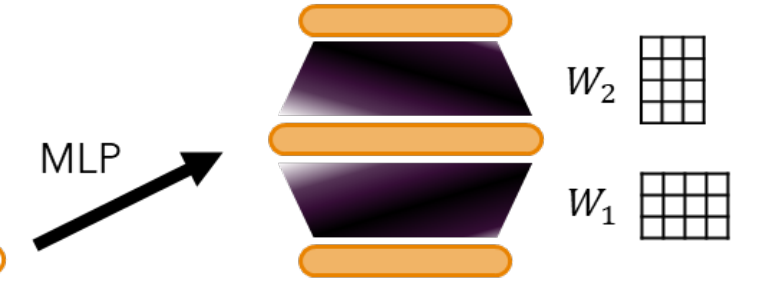


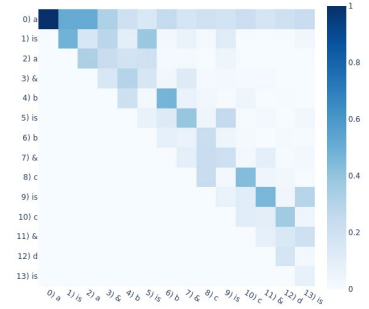
induction head
copies previous tokens

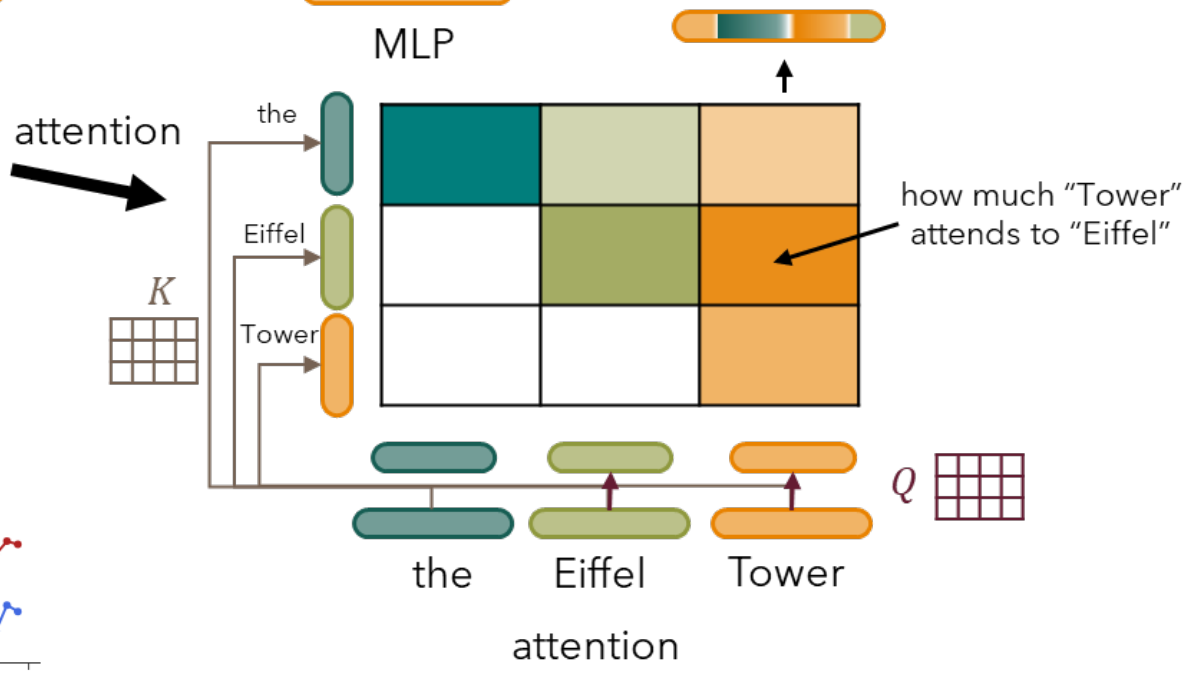
Summary

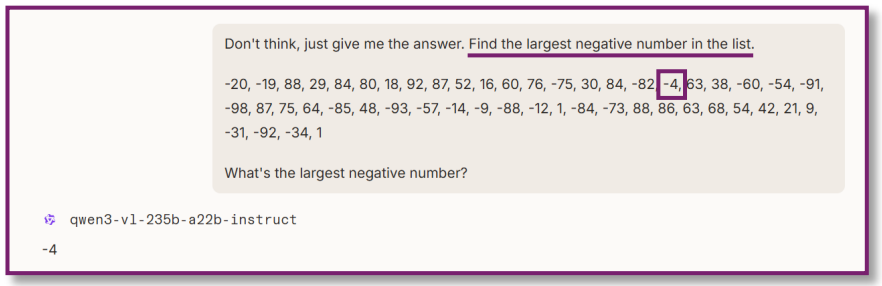

probability
token id
Paris
3
14
15
92
6
the
Eiffel
Tower
is
in

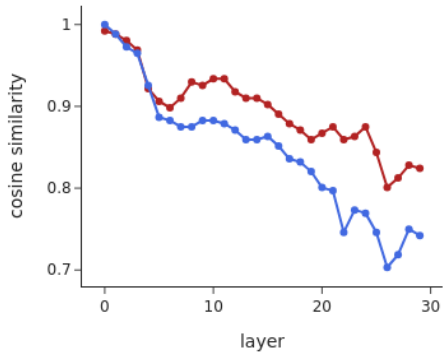

residual stream
Tower
...


MLP
W_2
W_1
MLP




attention
the
Eiffel
Tower
K
how much "Tower" attends to "Eiffel"
Q
the
Eiffel
Tower
attention


Don't think, just give me the answer. Find the largest negative number in the list.
-20, -19, 88, 29, 84, 80, 18, 92, 87, 52, 16, 60, 76, -75, 30, 84, -82, -4, 63, 38, -60, -54, -91, -98, 87, 75, 64, -85, 48, -93, -57, -14, -9, -88, -12, 1, -84, -73, 88, 86, 63, 68, 54, 42, 21, 9, -31, -92, -34, 1
What's the largest negative number?
qwen3-vl-235b-a22b-instruct
-4


cosine similarity
layer