Understanding Emergent Structure in Large Language Models

Ellie Pavlick, November 25, 2024

BROWN

L U N A R

Language Understanding and Representation Laboratory

The neuro-symbolic tug-of-war

ic tug-of-war

Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com

Noam Shazeer* Google Brain noam@google.com

Niki Parmar* Google Research nikip@google.com

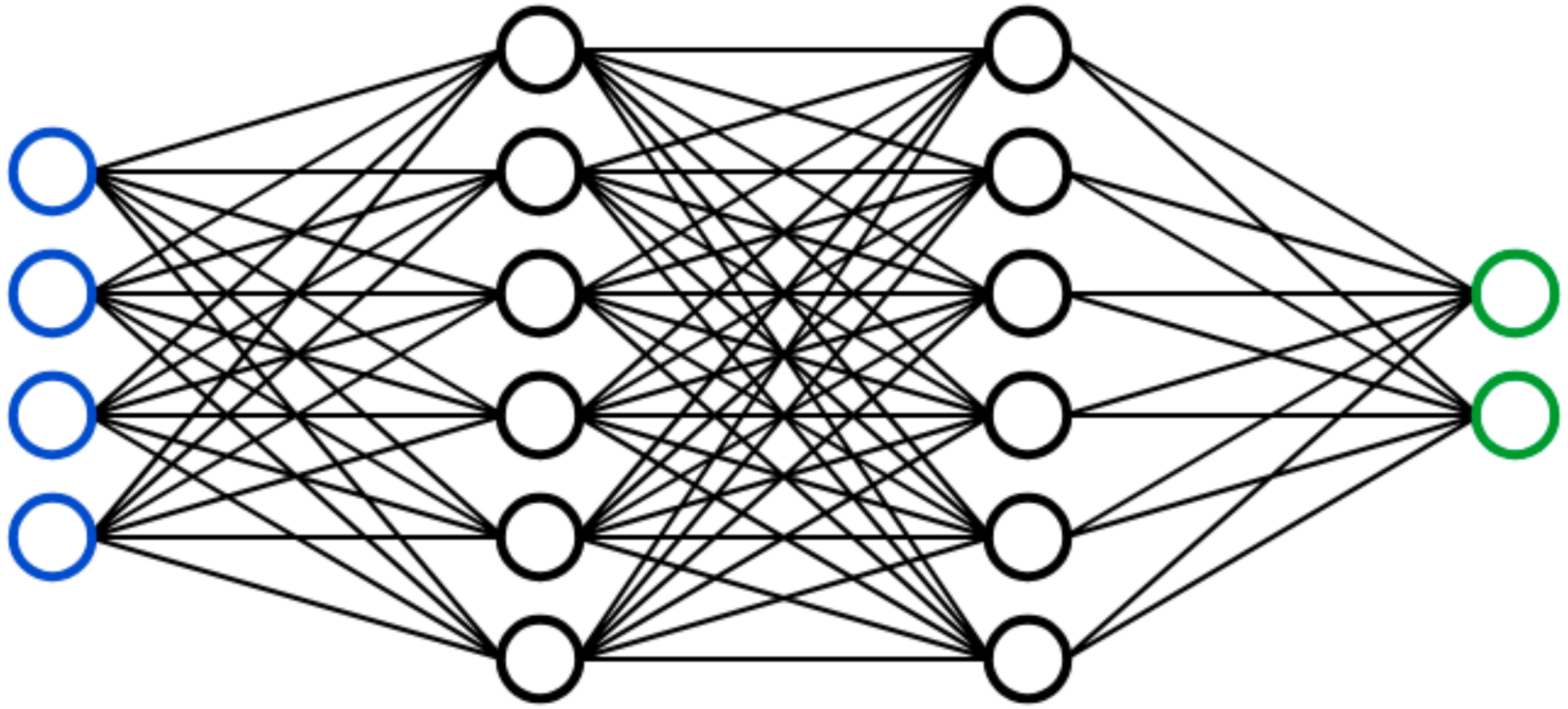
Jakob Uszkoreit* Google Research usz@google.com

Llion Jones* Google Research llion@google.com

Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu

Łukasz Kaiser* Google Brain lukaszkaiser@google.com

Illia Polosukhin* ‡ illia.polosukhin@gmail.com



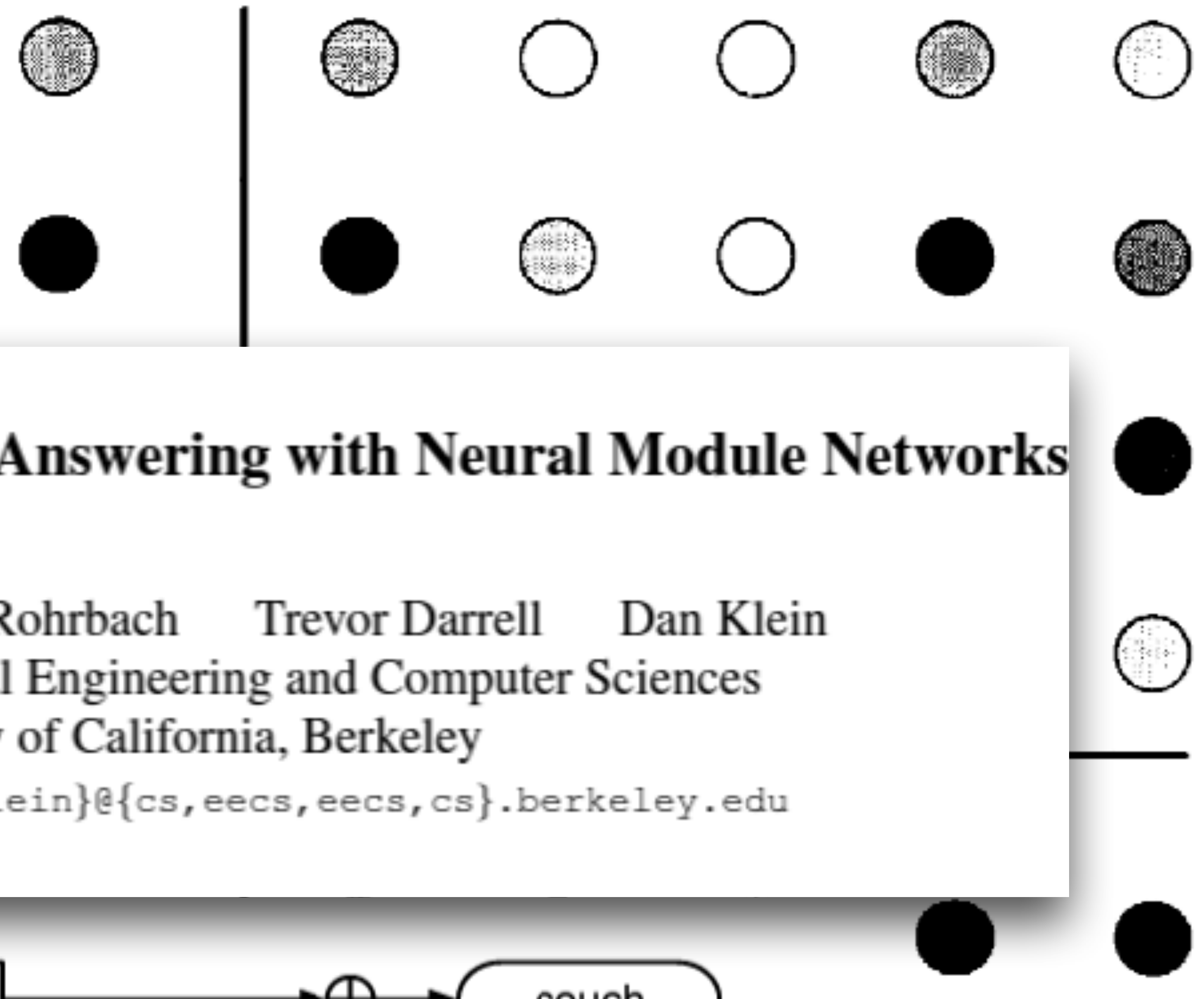
Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com
Noam Shazeer* Google Brain noam@google.com
Niki Parmar* Google Research nikip@google.com
Jakob Uszkoreit* Google Research usz@google.com
Llion Jones* Google Research llion@google.com
Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu
Łukasz Kaiser* Google Brain lukaszkaiser@google.com
Illia Polosukhin* ‡ illia.polosukhin@gmail.com

ic tug-of-

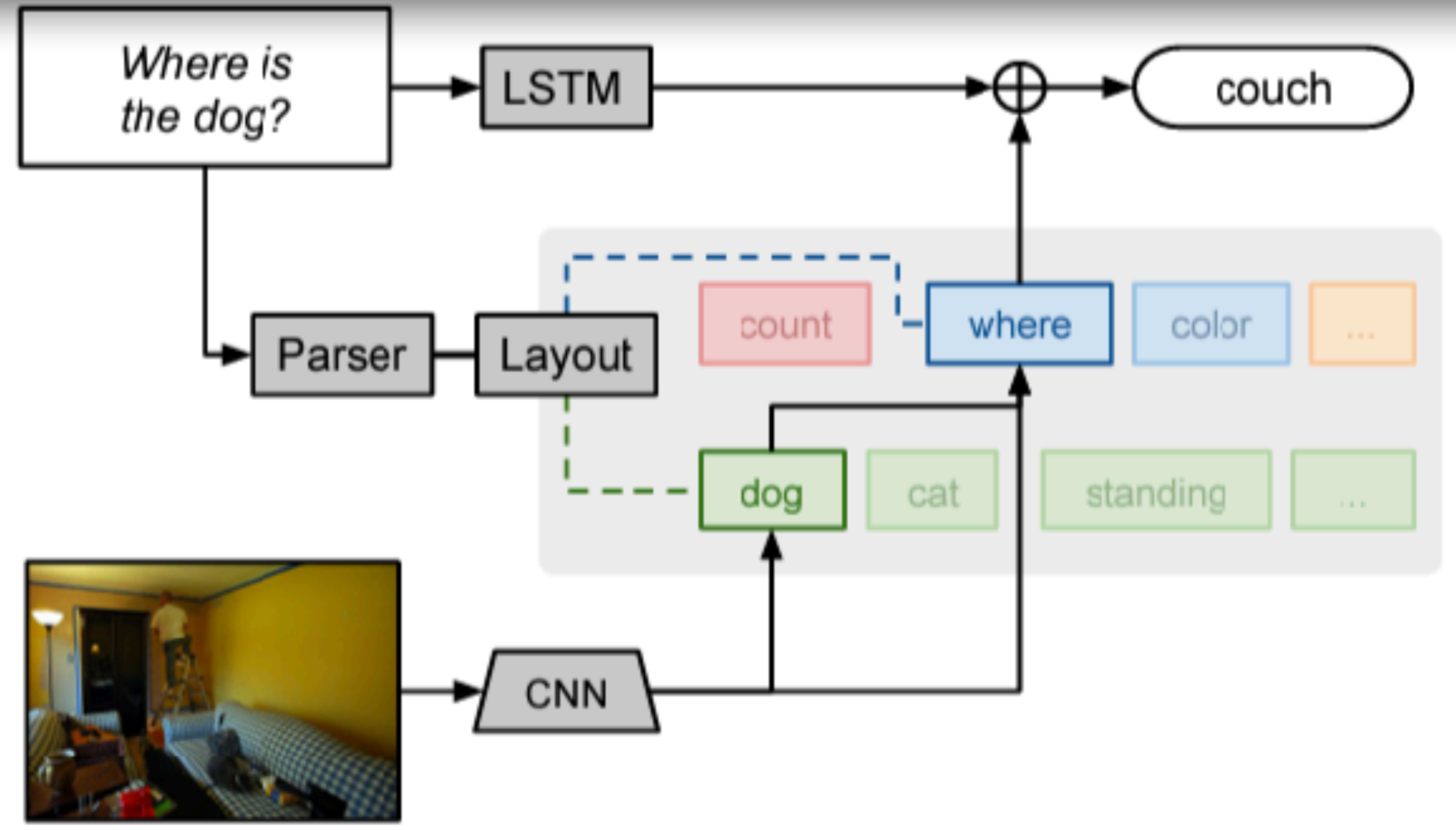
Tensor Product Variable Binding and the Representation of Symbolic Structures in Connectionist Systems

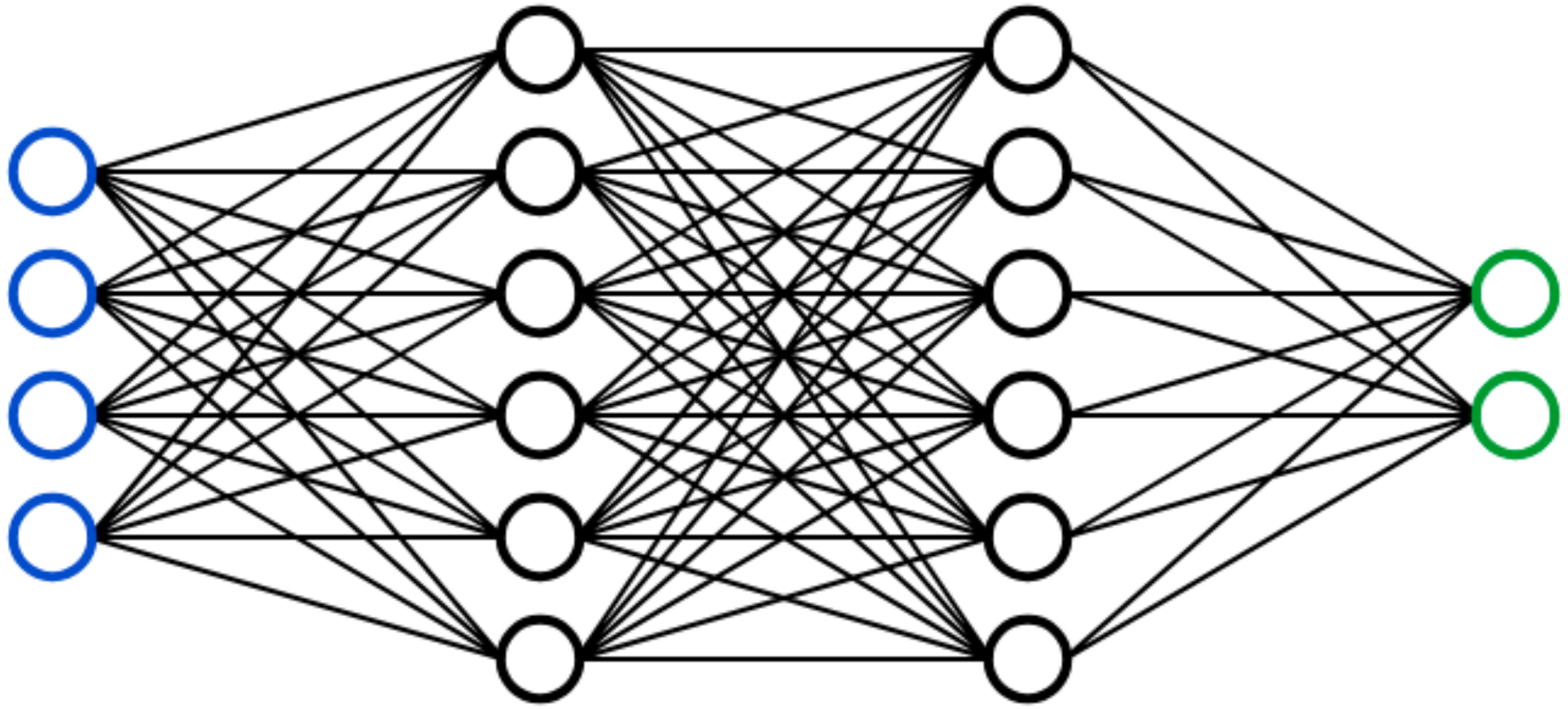
Paul Smolensky
Department of Computer Science and Institute of Cognitive Science, University of Colorado, Boulder, CO 80309-0430, USA



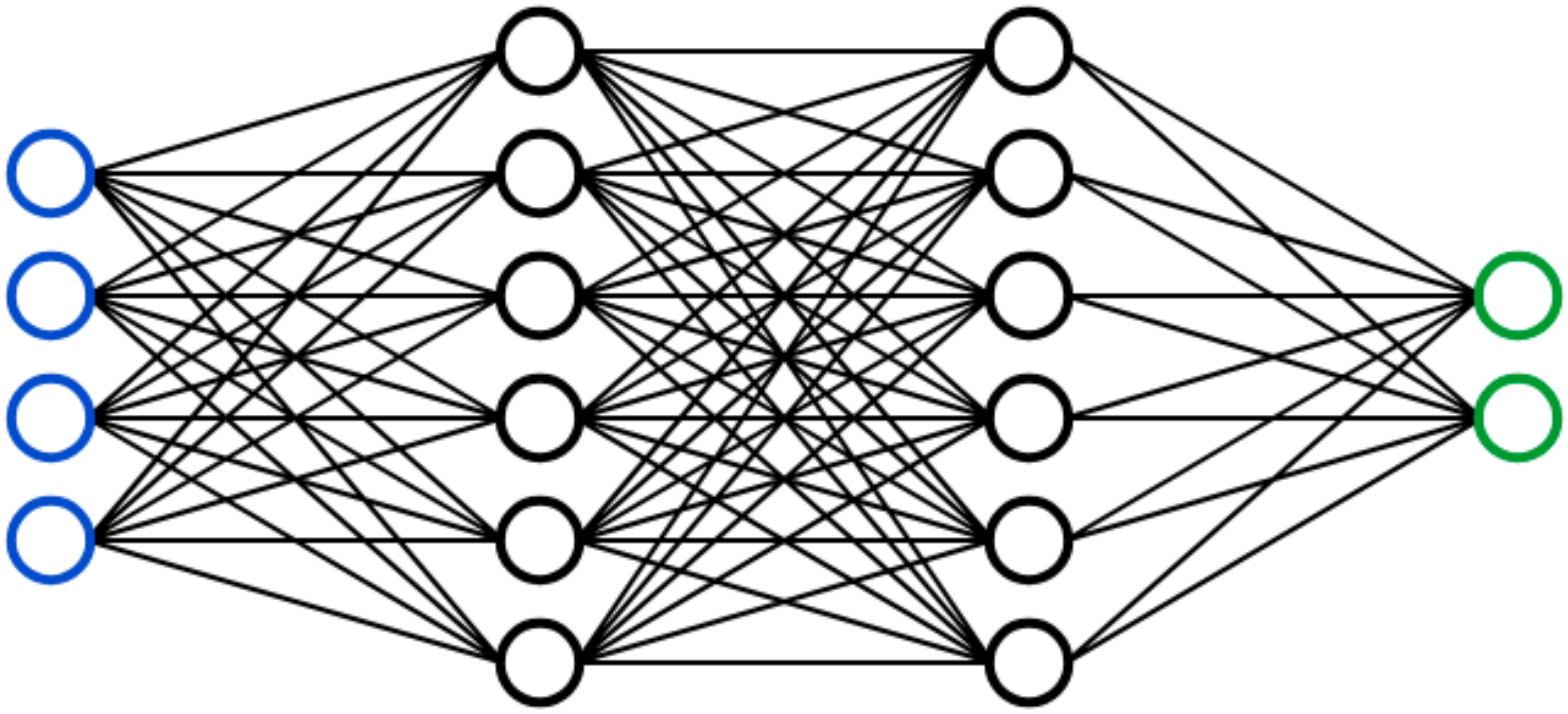
Deep Compositional Question Answering with Neural Module Networks

Jacob Andreas Marcus Rohrbach Trevor Darrell Dan Klein
Department of Electrical Engineering and Computer Sciences
University of California, Berkeley
{jda,rohrbach,trevor,klein}@{cs,eecs,eecs,cs}.berkeley.edu





Transformers aren't just webs of associations



Transformers aren't just webs of associations

Transformer Feed-Forward Layers Are Key-Value Memories

Mor Geva[1,2] Roei Schuster[1,3] Jonathan Berant[1,2] Omer Levy[1]
[1]Blavatnik School of Computer Science, Tel-Aviv University
[2]Allen Institute for Artificial Intelligence
[3]Cornell Tech
{morgeva@mail,joberant@cs,levyomer@cs}.tau.ac.il,rs864@cornell.edu

In-context Learning and Induction Heads

AUTHORS
Catherine Olsson*, Nelson Elhage*, Neel Nanda*, Nicholas Joseph†, Nova DasSarma†, Tom Henighan†, Ben Mann†, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, Chris Olah‡

AFFILIATION
Anthropic

PUBLISHED
Mar 8, 2022

* Core Research Contributor; † Core Infrastructure Contributor; ‡ Correspondence to colah@anthropic.com; Author contributions statement below.

INTERPRETABILITY IN THE WILD: A CIRCUIT FOR INDIRECT OBJECT IDENTIFICATION IN GPT-2 SMALL

Kevin Wang[1], Alexandre Variengien[1], Arthur Conmy[1], Buck Shlegeris[1] & Jacob Steinhardt[1,2]
[1]Redwood Research
[2]UC Berkeley
kevin@rdwrs.com, alexandre@rdwrs.com,
arthur@rdwrs.com, buck@rdwrs.com, jsteinhardt@berkeley.edu

Locating and Editing Factual Associations in GPT

Kevin Meng* MIT CSAIL
David Bau* Northeastern University
Alex Andonian MIT CSAIL
Yonatan Belinkov† Technion – IIT

Transformers aren't just webs of associations

Transformer Feed-Forward Layers Are Key-Value Memories

Mor Geva[1,2] Roei Schuster[1,3] Jonathan Berant[1,2] Omer Levy[1]
[1]Blavatnik School of Computer Science, Tel-Aviv University

Key-Value Stores

In-context Learning and Induction Heads

Read-Write Memory/Registers

* Core Research Contributor; † Core Infrastructure Contributor; ✱ Correspondence to colah@anthropic.com;
Author contributions statement below.

INTERPRETABILITY IN THE WILD: A CIRCUIT FOR INDIRECT OBJECT IDENTIFICATION IN GPT-2 SMALL

Interpretable algorithms playing out over layers

Locating and Editing Factual Associations in GPT

Mutable "Knowledge Bases"

Why care about what’s inside the black box?

1. Curiosity :)
2. Safety — Understanding the “source code” can help us anticipate when and how things might go wrong
3. Theory — Boiling LLMs down into computational building blocks might enable us to develop more principled mathematical theories of representations and learning
4. Engineering — Knowing how things work could allow us to acheive the same results more quickly, reliably, cheaply
5. Cognitive, Linguistic, Neuro- Science — AI could serve as a source of new hypotheses and theories about the nature of language and cognition in general

This Talk

- Transformers and the “Mental Model of LLMs”
- Two Proofs of Concept:
    - Abstract representation of relations
    - Modular and reusable algorithmic “building blocks”

This Talk

- **Transformers and the “Mental Model of LLMs”**
- Two Proofs of Concept:
  - Abstract representation of relations
  - Modular and reusable algorithmic “building blocks”

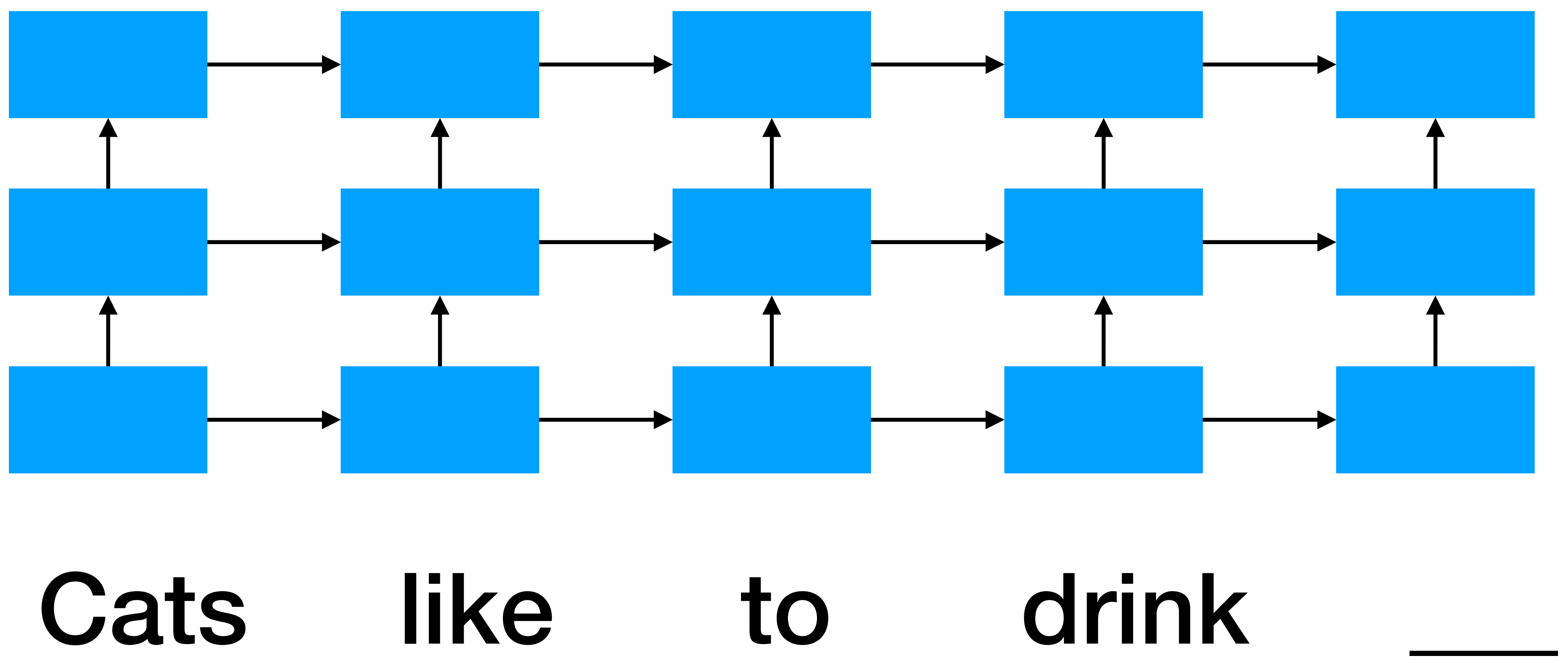
Mental model of LLMs

Neural Nets for Sequence Modeling

Mental model of LLMs

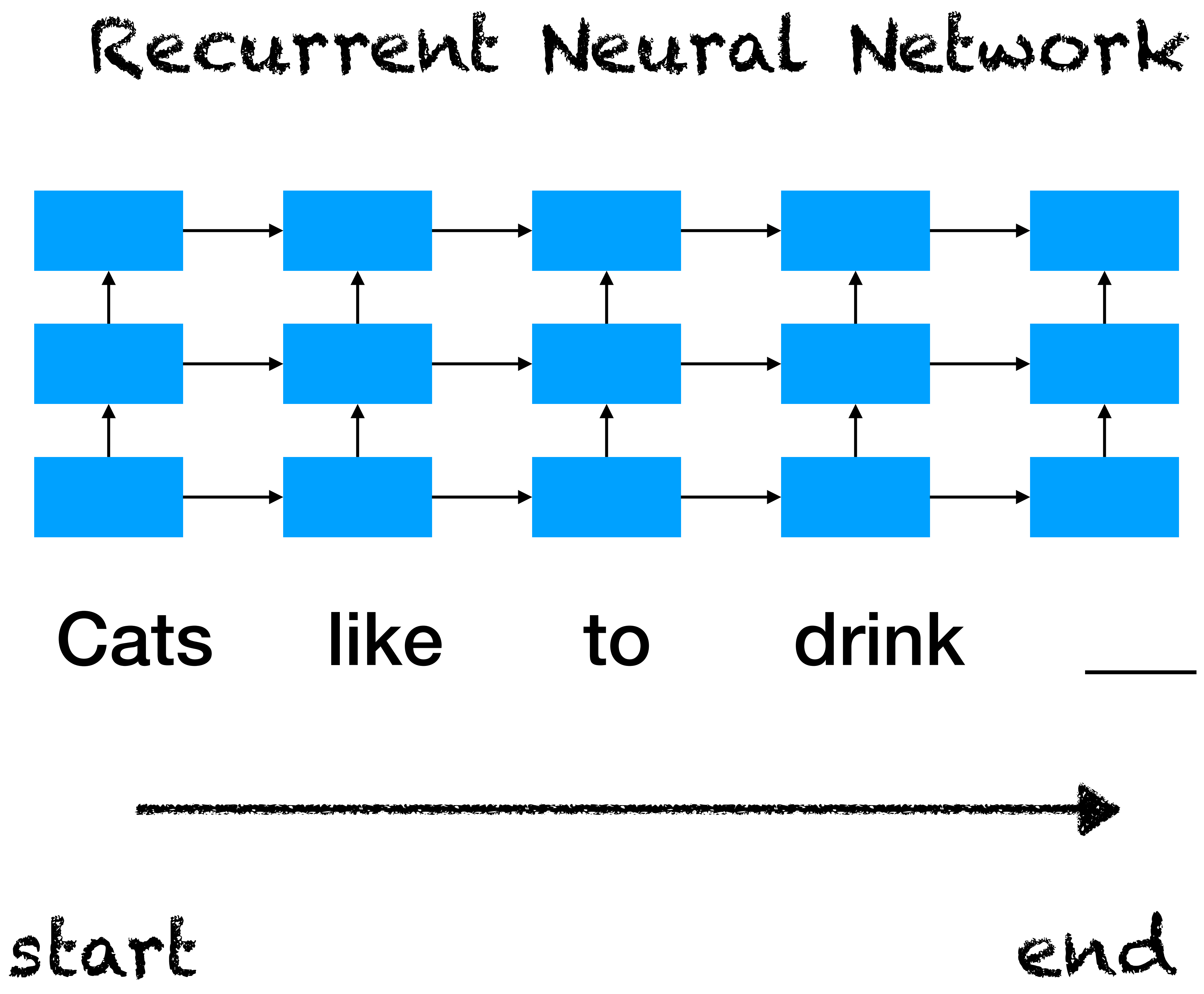
Neural Nets for Sequence Modeling

Recurrent Neural Network



Mental model of LLMs

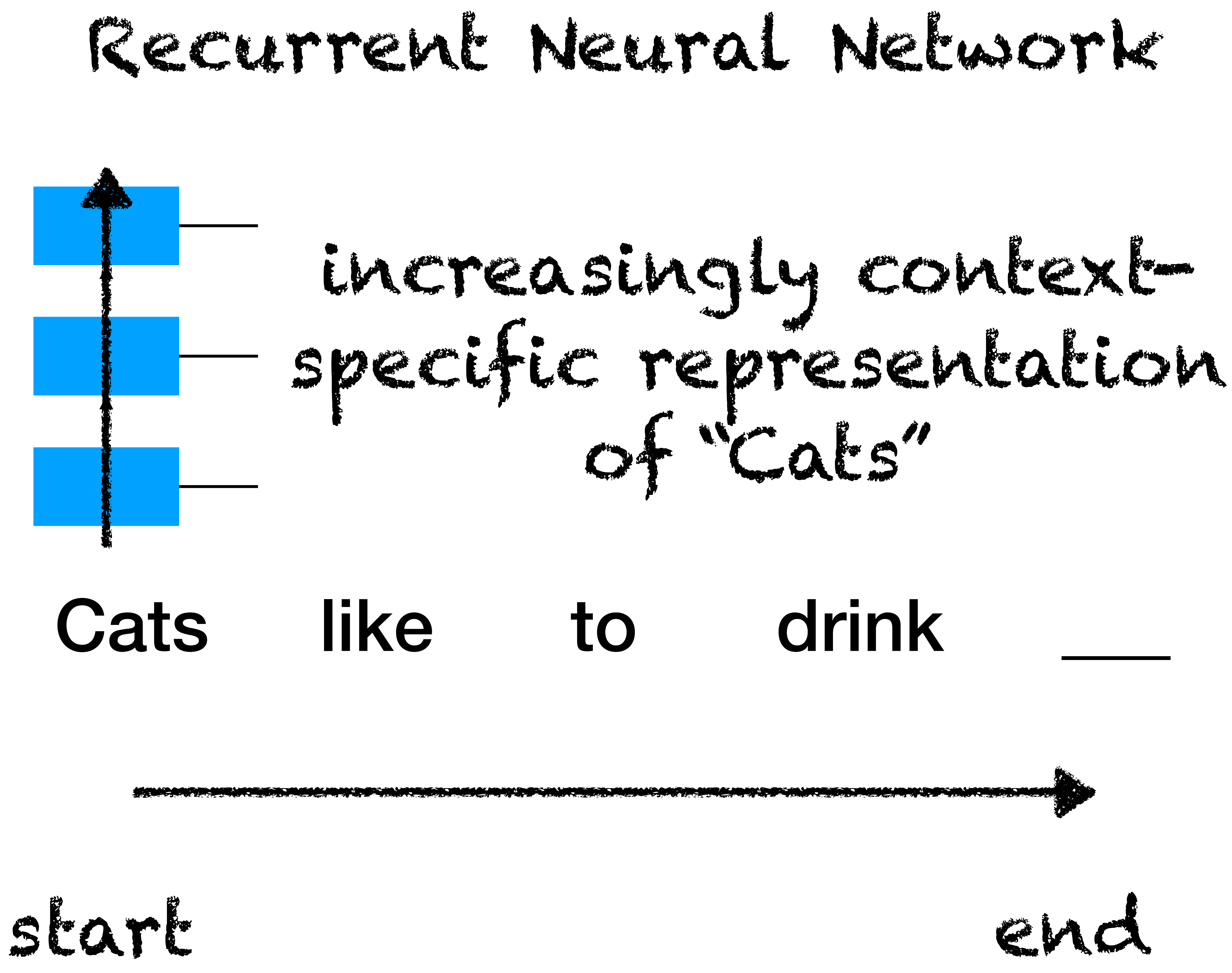
Neural Nets for Sequence Modeling



Assumption #1:
(Compute) time goes
left to right

Mental model of LLMs

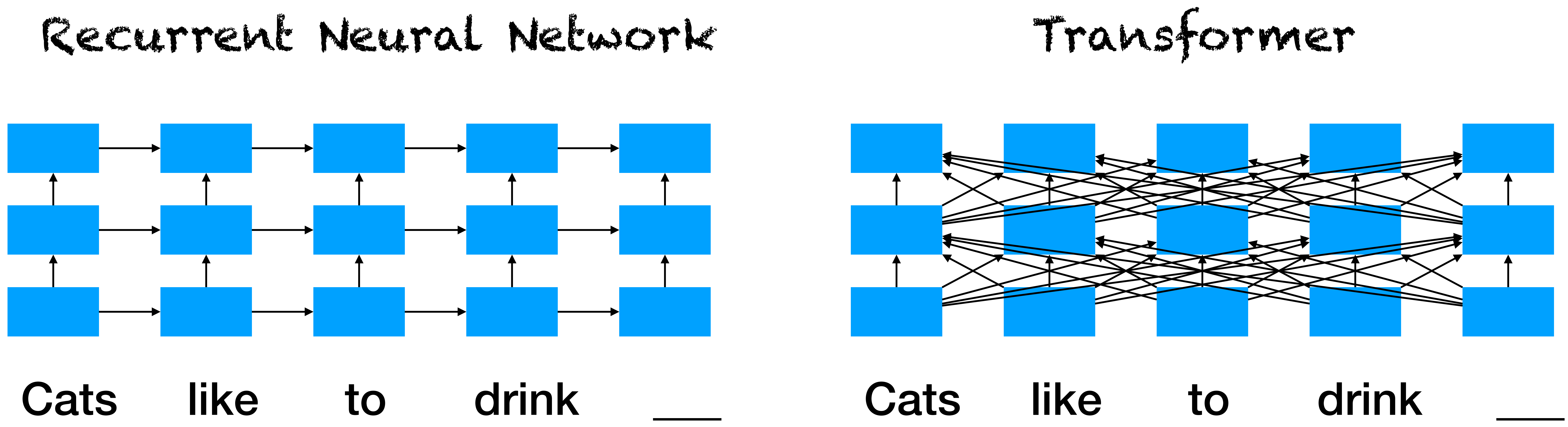
Neural Nets for Sequence Modeling



Assumption #2: Token embeddings represent words

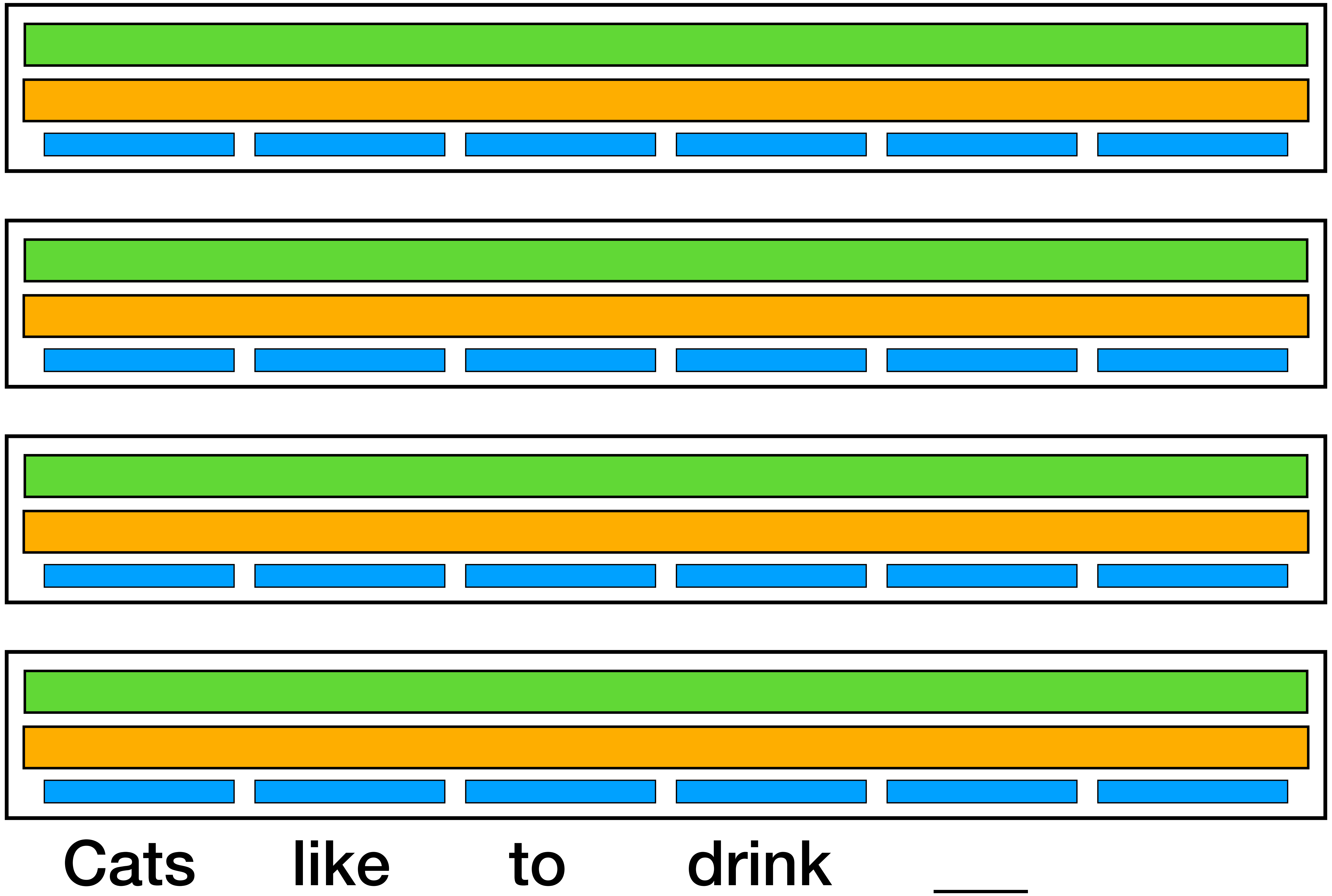
Mental model of LLMs

Neural Nets for Sequence Modeling



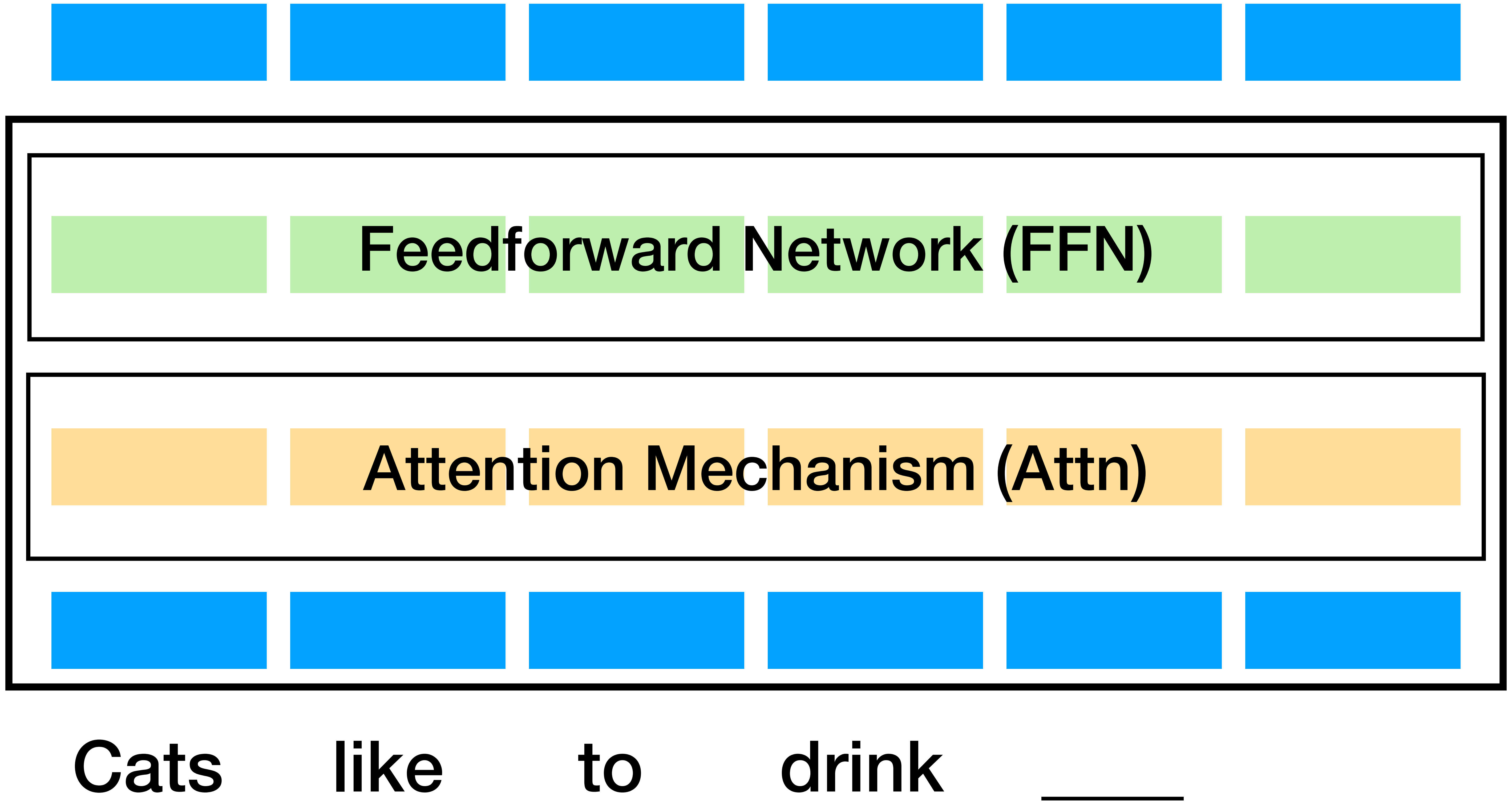
Mental model of LLMs

Transformer Architecture



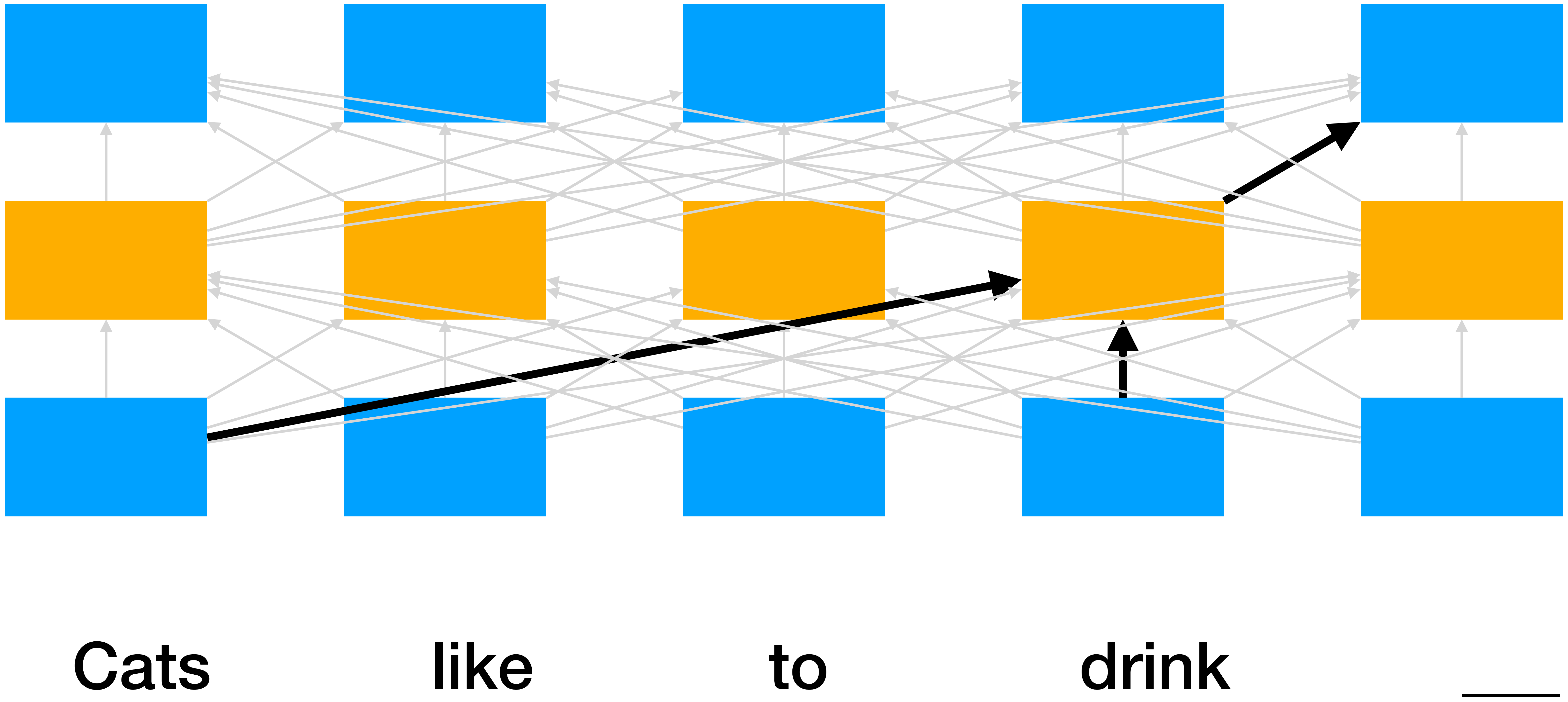
Mental model of LLMs

Transformer Architecture



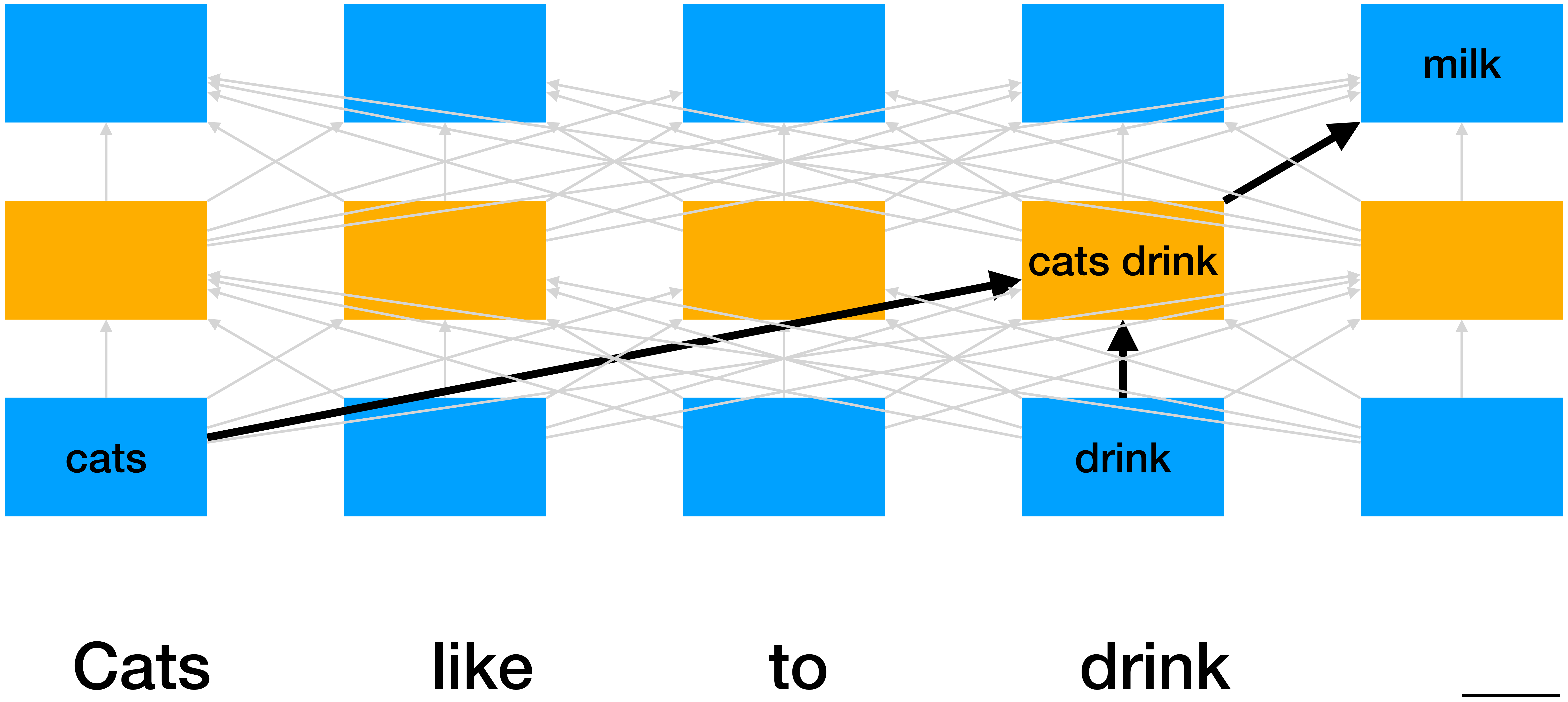
Mental model of LLMs

Transformer Architecture



Mental model of LLMs

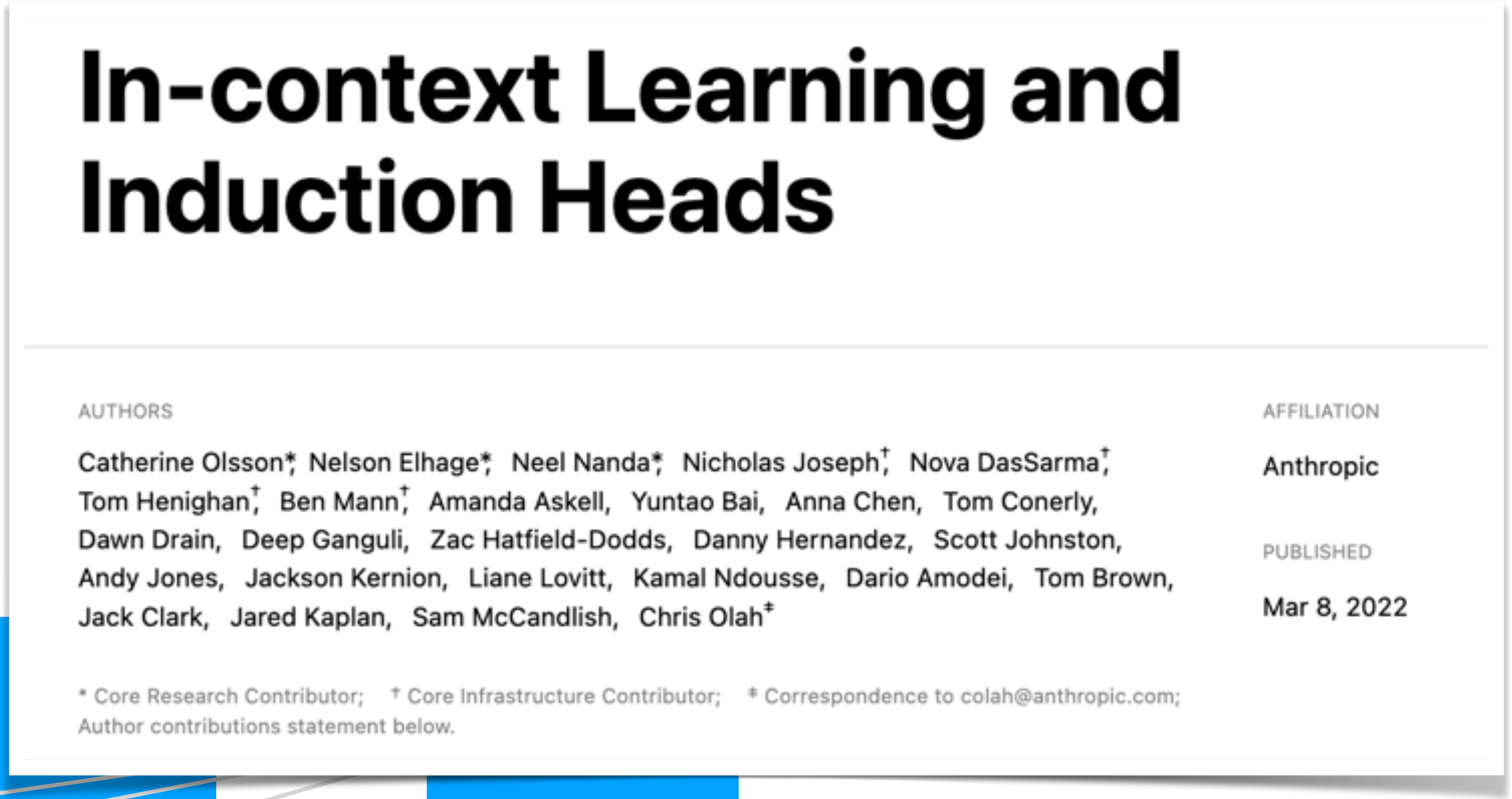
Transformer Architecture



Attention is a read-write mechanism. It reads from registers at one layer, and writes to registers in the next layer.

Mental model of LLMs

Transformer Architecture



In-context Learning and Induction Heads

AUTHORS

Catherine Olsson*, Nelson Elhage*, Neel Nanda*, Nicholas Joseph†, Nova DasSarma†, Tom Henighan†, Ben Mann†, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, Chris Olah‡

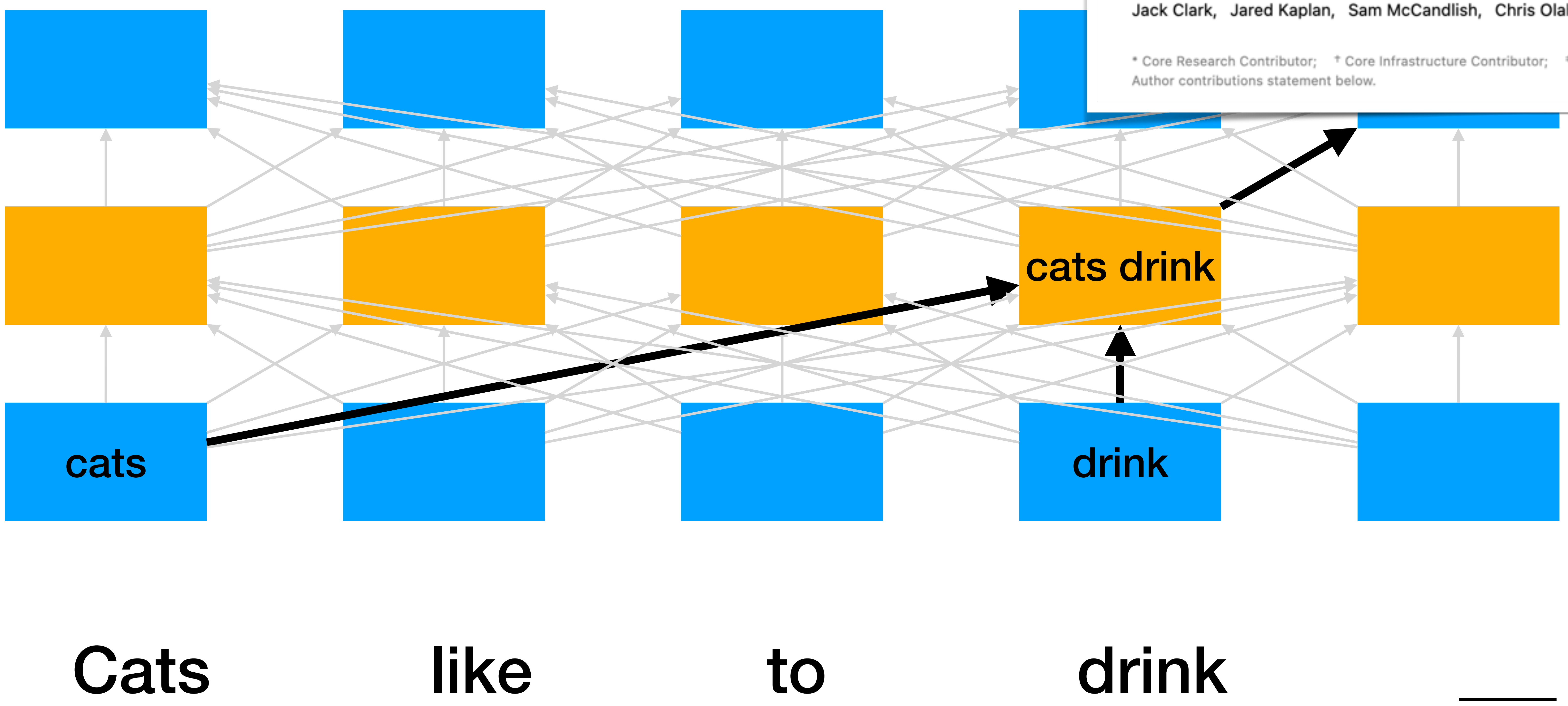
AFFILIATION

Anthropic

PUBLISHED

Mar 8, 2022

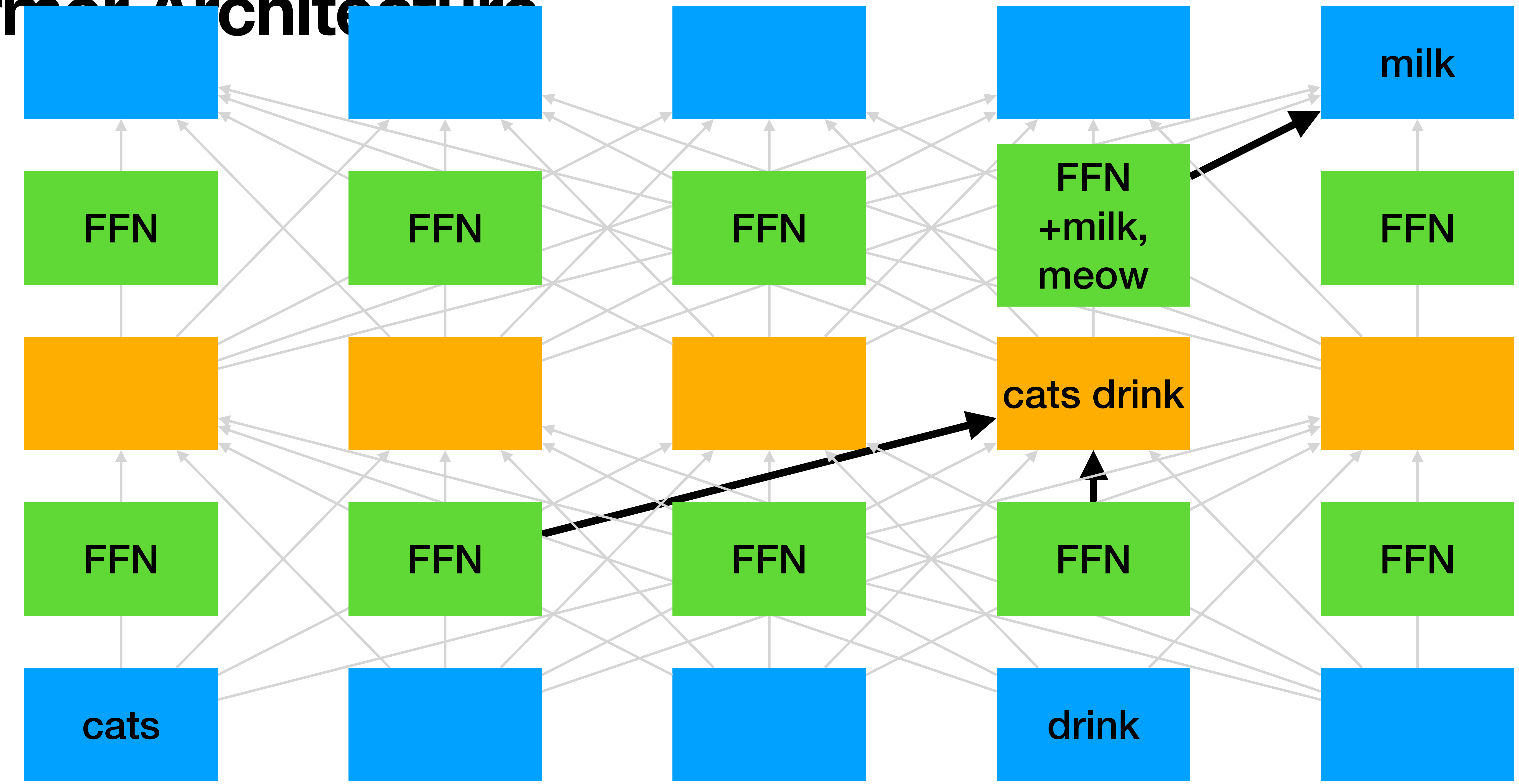
* Core Research Contributor; † Core Infrastructure Contributor; ‡ Correspondence to colah@anthropic.com; Author contributions statement below.



Attention is a read-write mechanism. It reads from registers at one layer, and writes to registers in the next layer.

Mental model of LLMs

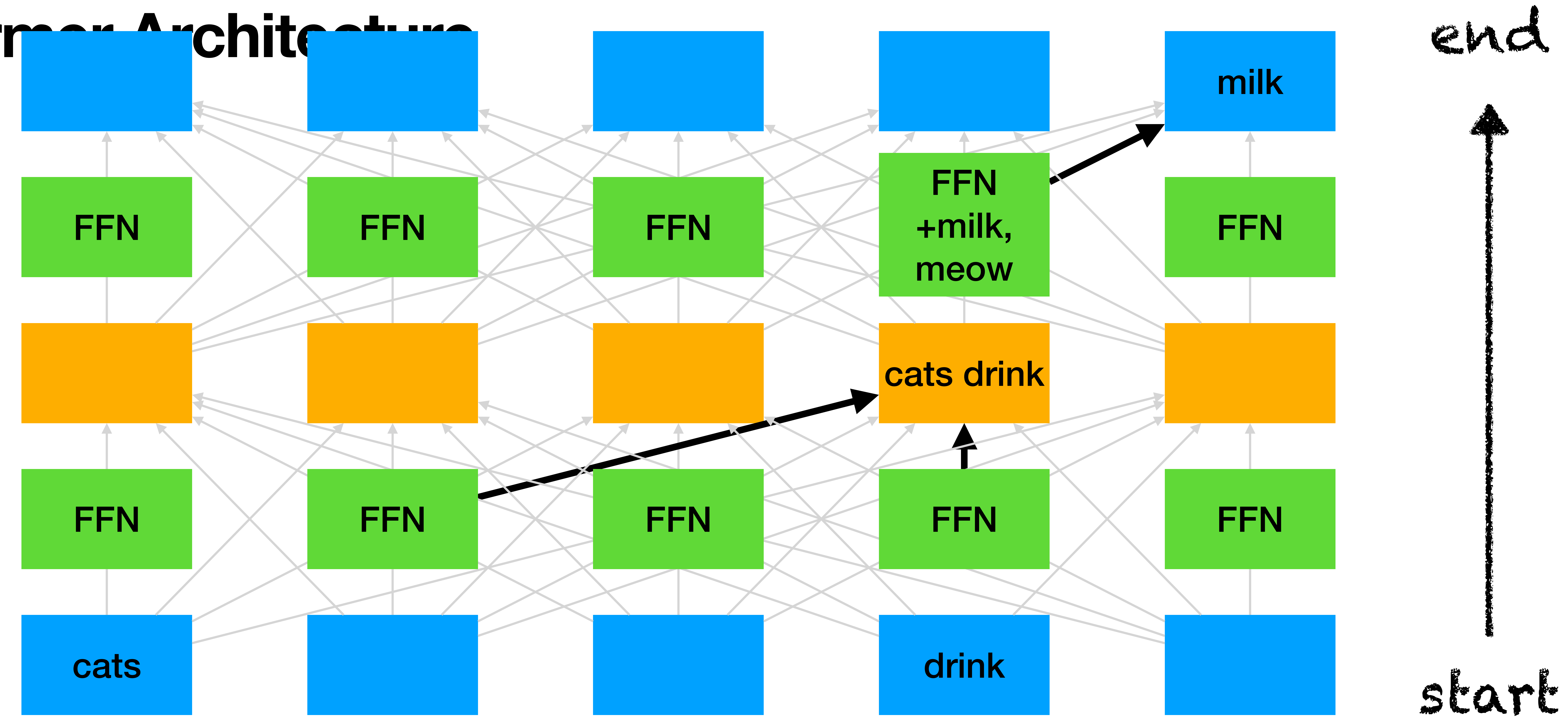
Transformer Architecture



Feed forward nets pull in new "stuff". I.e., add info into the registers based on recall from training.

Mental model of LLMs

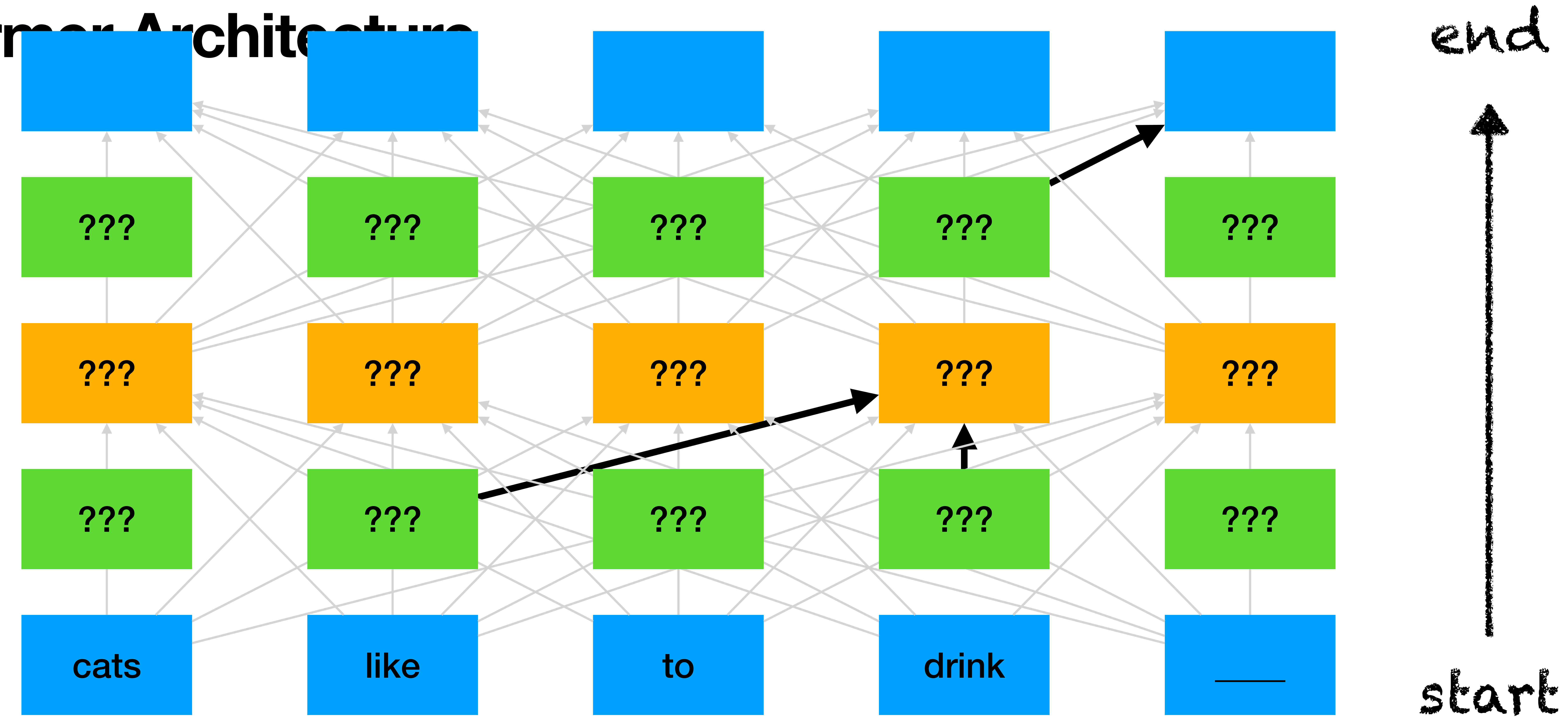
Transformer Architecture



(Compute) time goes bottom to top

Mental model of LLMs

Transformer Architecture



Register content can, in theory, be anything!

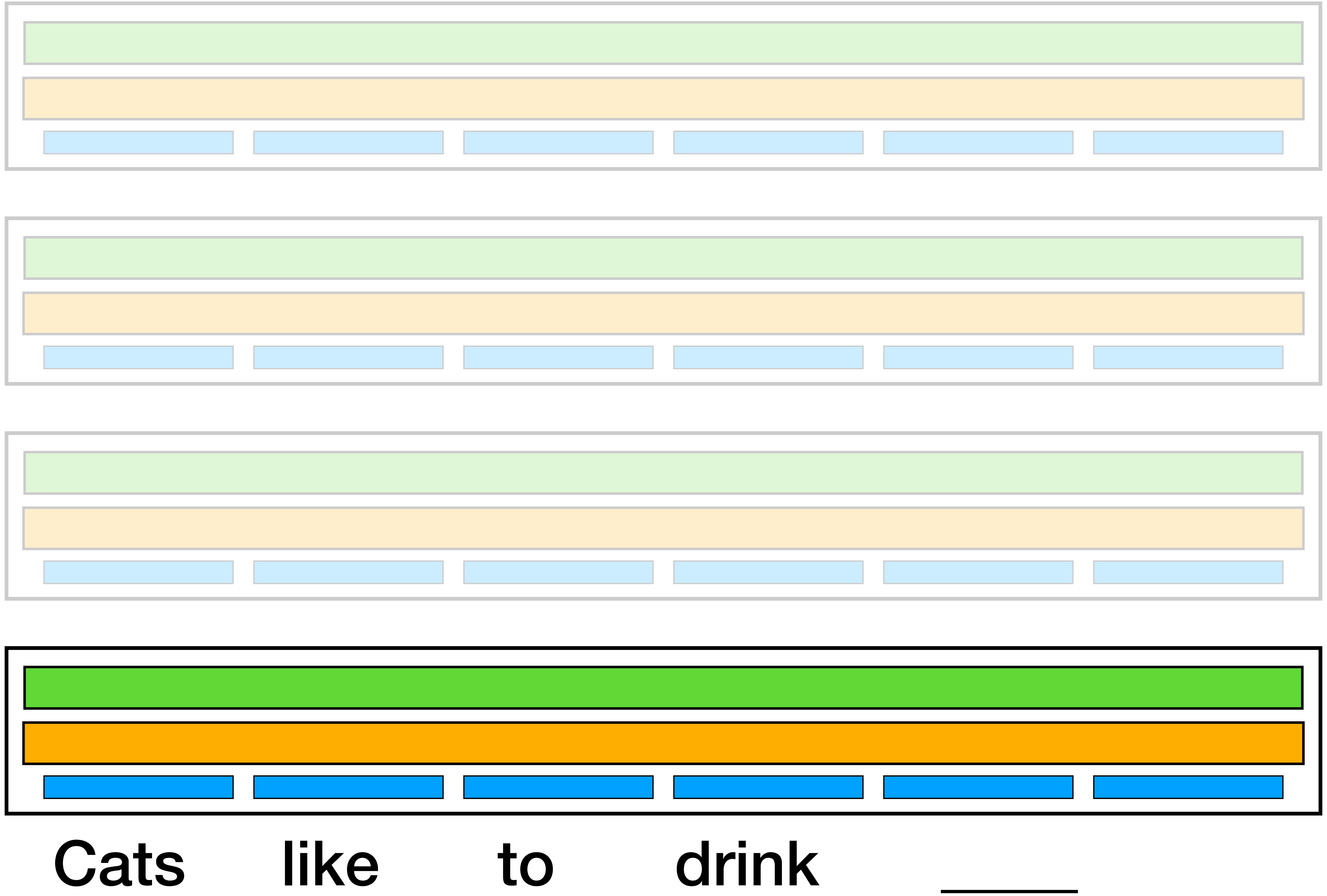
Mental model of LLMs

Transformer Feed-Forward Layers Build Predictions by Promoting Concepts in the Vocabulary Space

Mor Geva[*,1] Avi Caciularu[*,2,†] Kevin Ro Wang[3] Yoav Goldberg[1,2]
[1]Allen Institute for AI [2]Bar-Ilan University [3]Independent Researcher
morp@allenai.org,{avi.c33,kevinrowang,yoav.goldberg}@gmail.com

Each layer makes an intermediate update to the predicted next token in vocab space. This "residual stream" is the input to the next layer.

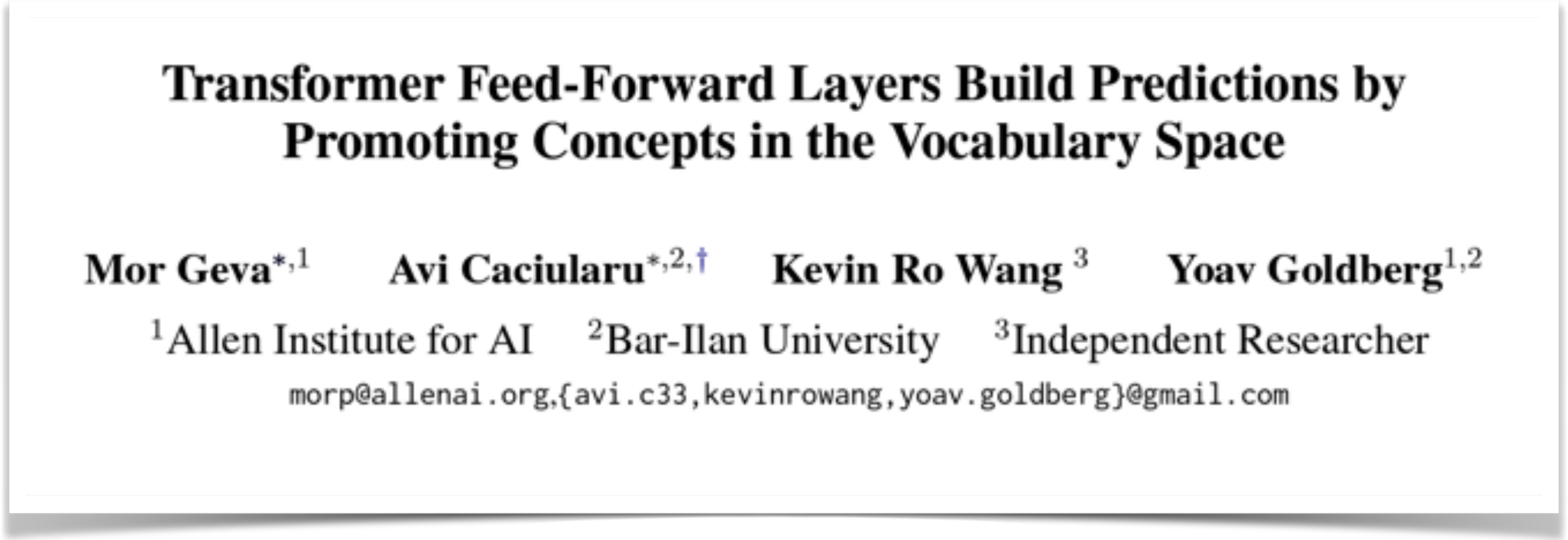
Architecture



Residual Stream

the
a
of
.
(
#
<html>
.
.
.

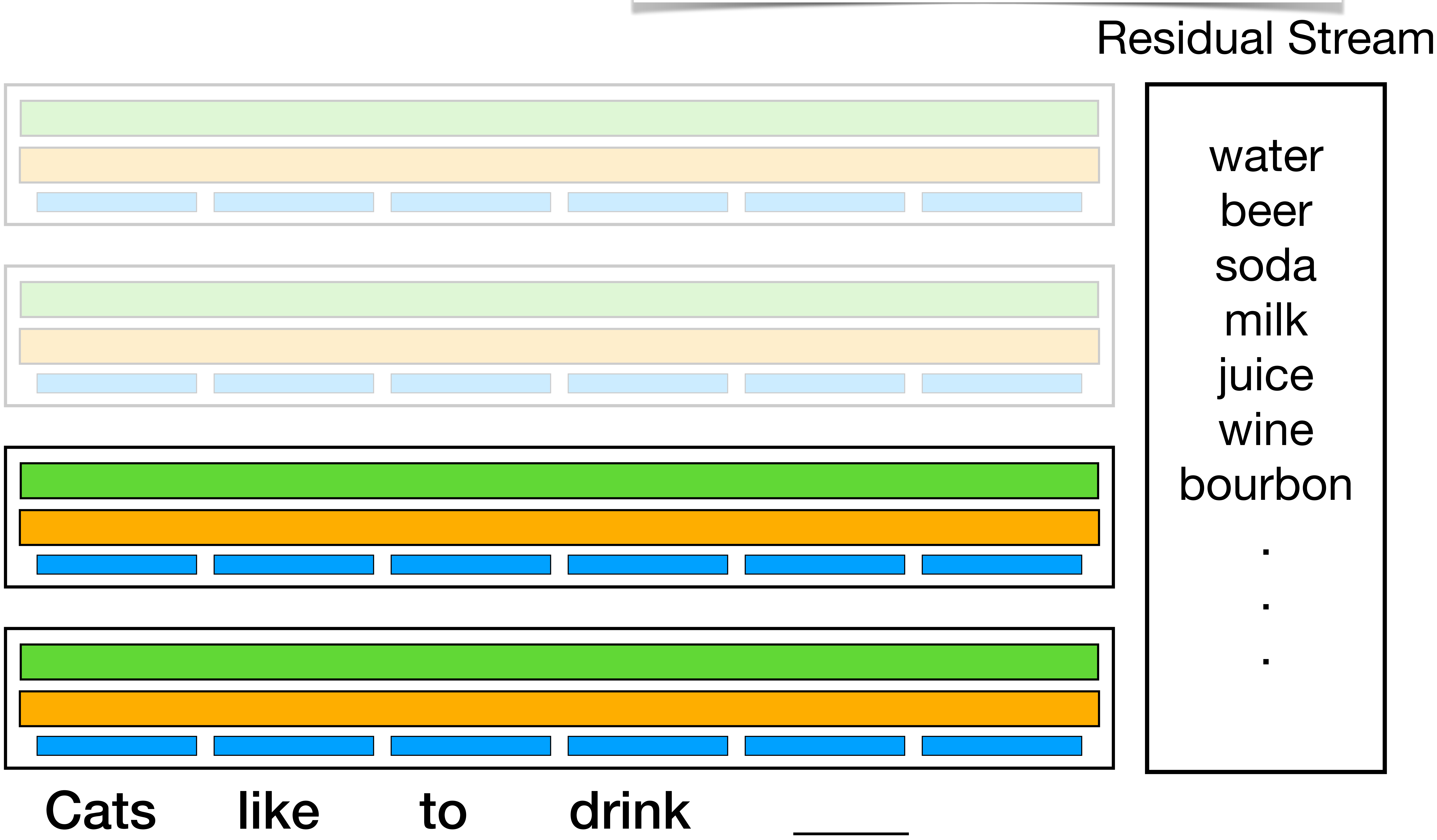
Mental model of LLMs


Transformer Feed-Forward Layers Build Predictions by Promoting Concepts in the Vocabulary Space

Mor Geva*,[1] Avi Caciularu*,[2,†] Kevin Ro Wang[3] Yoav Goldberg[1,2]
[1]Allen Institute for AI [2]Bar-Ilan University [3]Independent Researcher
morp@allenai.org,{avi.c33,kevinrowang,yoav.goldberg}@gmail.com

Each layer makes an intermediate update to the predicted next token in vocab space. This "residual stream" is the input to the next layer.

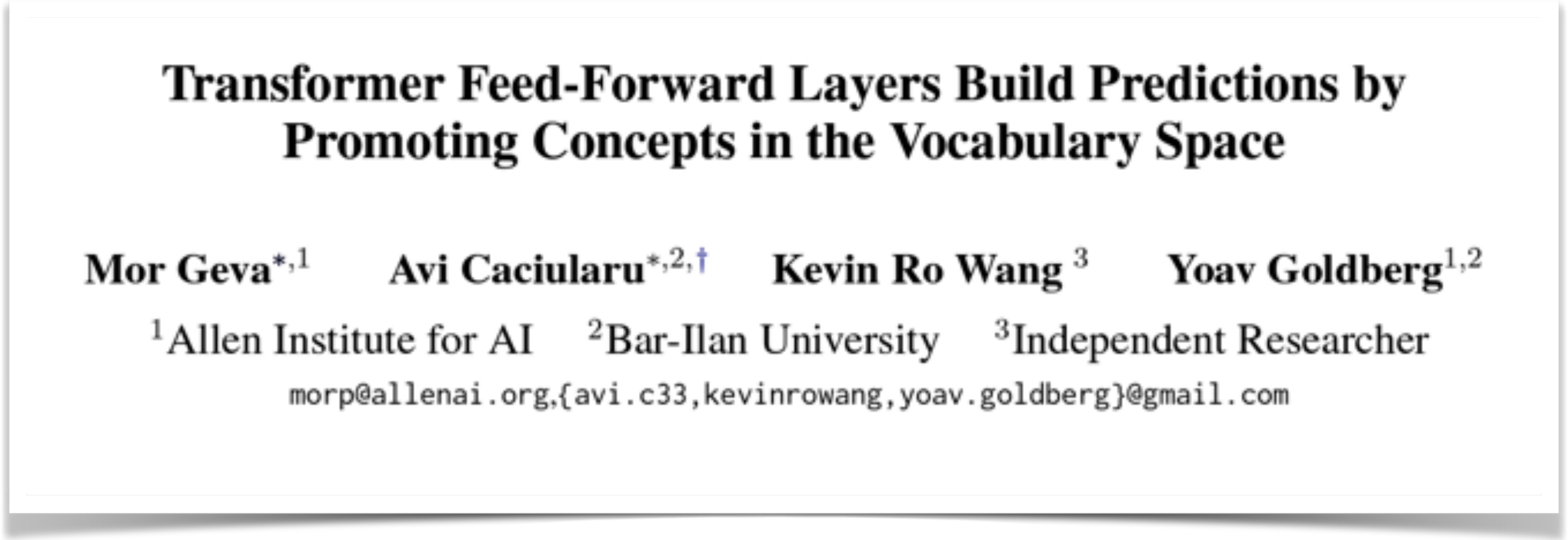
Architecture



Mental model of LLMs

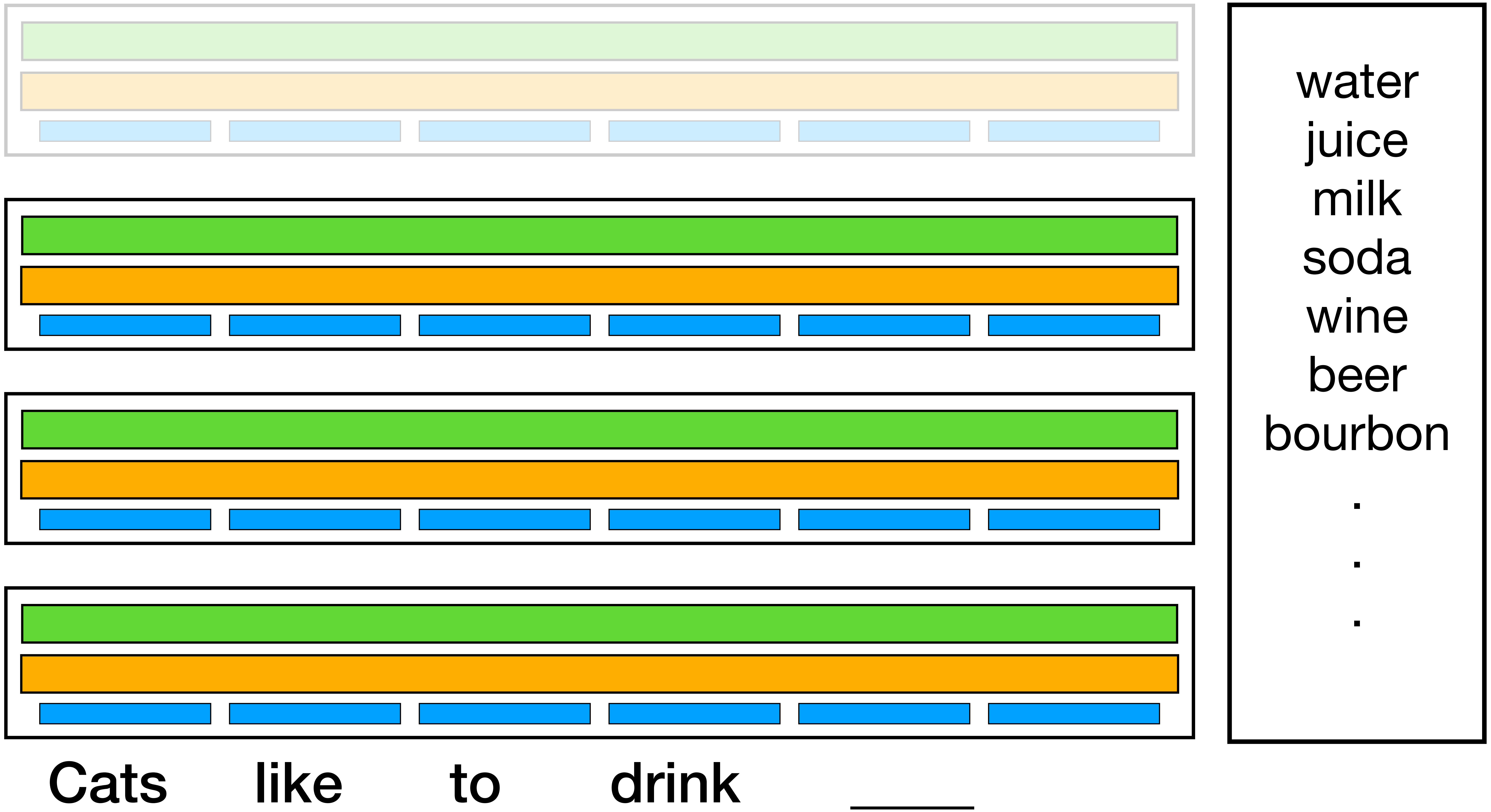
Each layer makes an intermediate update to the predicted next token in vocab space. This "residual stream" is the input to the next layer.

Architecture

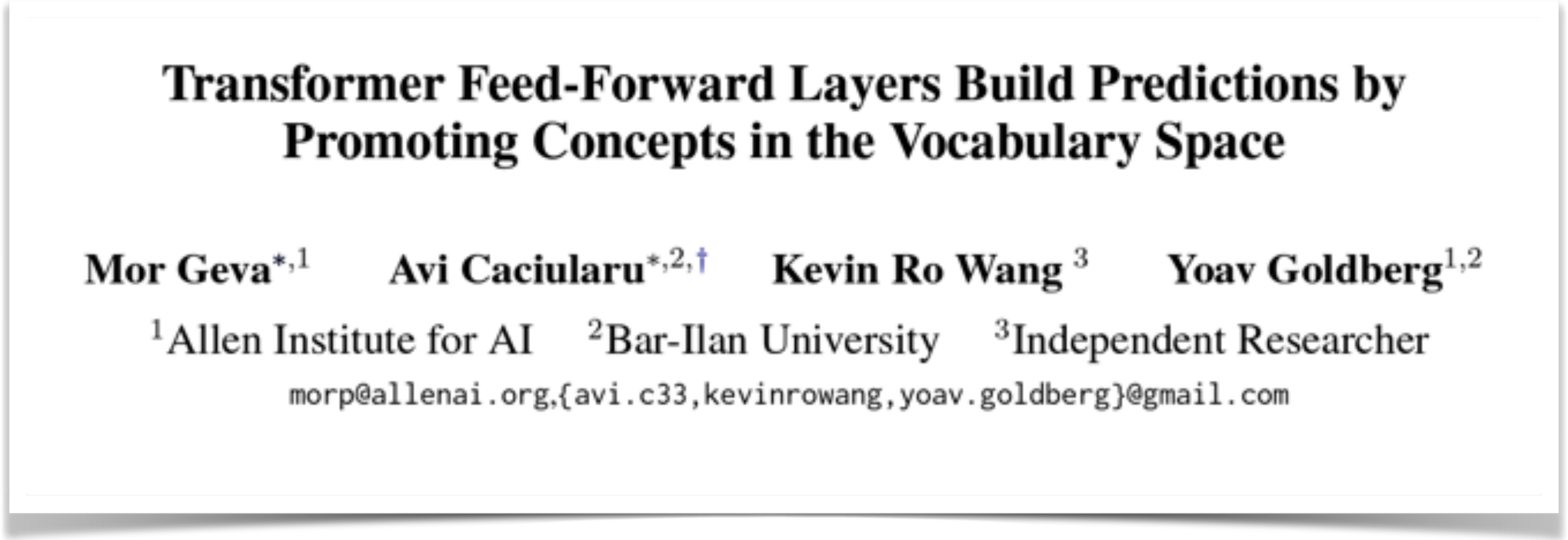

Transformer Feed-Forward Layers Build Predictions by Promoting Concepts in the Vocabulary Space

Mor Geva*,[1] Avi Caciularu*,[2,†] Kevin Ro Wang[3] Yoav Goldberg[1,2]
[1]Allen Institute for AI [2]Bar-Ilan University [3]Independent Researcher
morp@allenai.org,{avi.c33,kevinrowang,yoav.goldberg}@gmail.com

Residual Stream



Mental model of LLMs


Transformer Feed-Forward Layers Build Predictions by Promoting Concepts in the Vocabulary Space

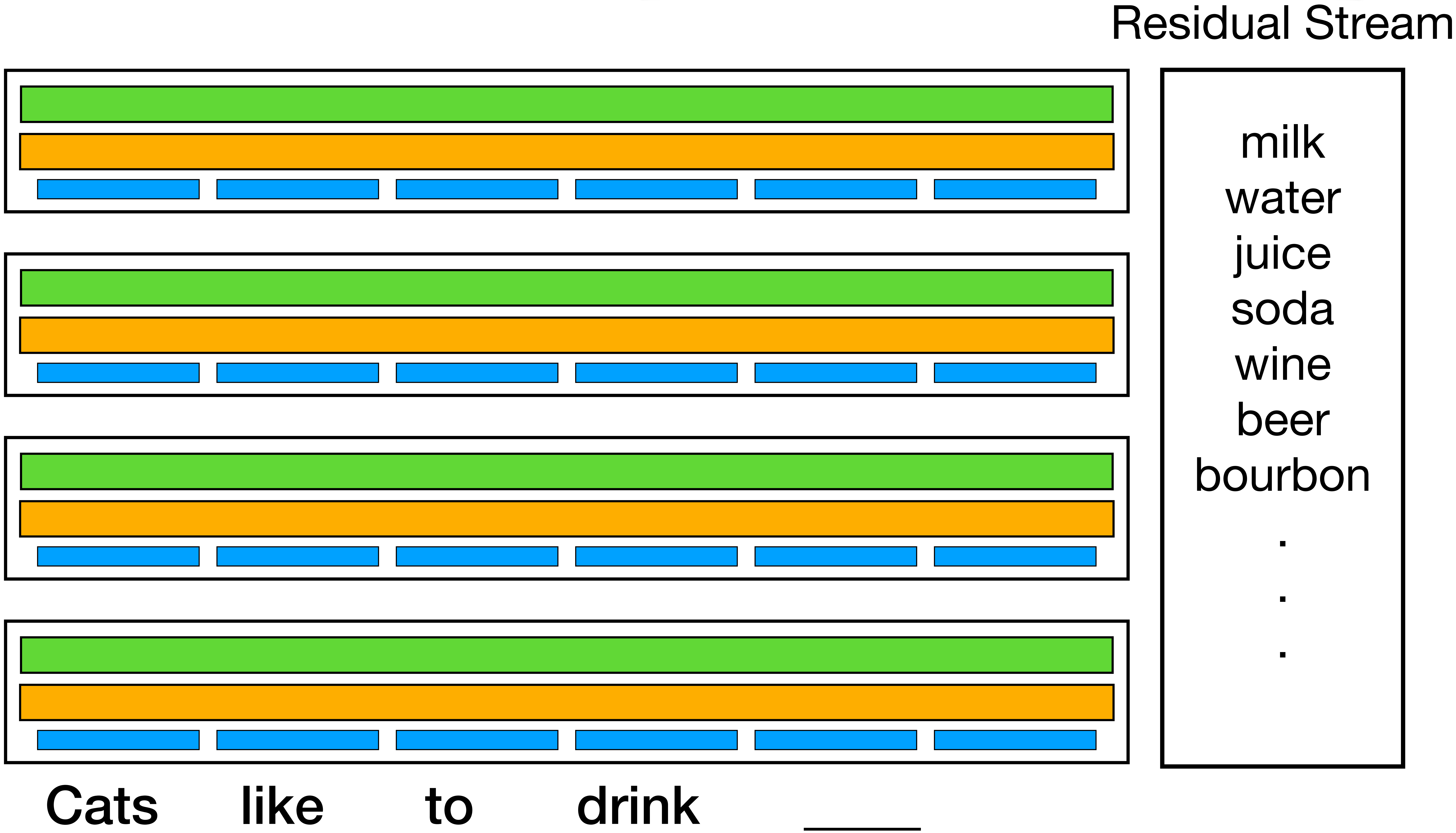
Mor Geva[*,1] Avi Caciularu[*,2,†] Kevin Ro Wang[3] Yoav Goldberg[1,2]

[1]Allen Institute for AI [2]Bar-Ilan University [3]Independent Researcher

morp@allenai.org,{avi.c33,kevinrowang,yoav.goldberg}@gmail.com

Each layer makes an intermediate update to the predicted next token in vocab space. This "residual stream" is the input to the next layer.

Architecture



Mental model of LLMs

Transformer Architecture Takeaways

- Attention Heads carry out reads-and-writes across layers. Tokens can be viewed as arbitrary “registers”.
- FFNs pull in new information from training (stuff not in local context).
- At each layer, we can get a kind of “print statement” showing the effect of these intermediate computations by looking at the effect on the residual stream

This Talk

- Transformers and the “Mental Model of LLMs”
- **Two Proofs of Concept:**
  - **Abstract representation of relations**
  - Modular and reusable algorithmic “building blocks”

Abstract Functions in LLMs

Jack Merullo

Qinan Yu

Carsten Eickhoff

Language Models Implement Simple Word2Vec-style Vector Arithmetic

Jack Merullo
Department of Computer Science
Brown University
jack_merullo@brown.edu

Carsten Eickhoff
School of Medicine
University of Tübingen
carsten.eickhoff@uni-tuebingen.de

Ellie Pavlick
Department of Computer Science
Brown University
ellie_pavlick@brown.edu

Characterizing Mechanisms for Factual Recall in Language Models

Qinan Yu **Jack Merullo** **Ellie Pavlick**
Brown University
Department of Computer Science
{qinan_yu,jack_merullo,ellie_pavlick}@brown.edu

Abstract Functions in LLMs

In-Context Learning Creates Task Vectors

Roee Hendel
Tel Aviv University
roee.hendel@mail.tau.ac.il

Mor Geva
Google DeepMind
pipek@google.com

Amir Globerson
Tel Aviv University, Google
gamir@tauex.tau.ac.il

Word2Vec-style Vector Arithmetic

Carsten Eickhoff
School of Medicine
University of Tübingen
rsten.eickhoff@uni-tuebingen.de

lick
nputer Science
versity
@brown.edu

FUNCTION VECTORS IN LARGE LANGUAGE MODELS

Eric Todd,* Millicent L. Li, Arnab Sen Sharma, Aaron Mueller,
Byron C. Wallace, and David Bau
Khoury College of Computer Sciences, Northeastern University

Carsten Eickhoff

Abstract Functions in LLMs

Task Setup

What is the capital of France?

Paris

What is the capital of Poland?

Warsaw

Abstract Functions in LLMs

Possible Mechanisms

Possibility #1: Models use idiomatic word associations to determine the probability of the next word.

What is the capital of France?

Paris

What is the capital of Poland?

Warsaw

P(Warsaw | Poland & of & capital & … of Poland & capital of & …)

Abstract Functions in LLMs

Possible Mechanisms

Possibility #2: Models infer an abstract function based on example, and then apply it to the input.

What is the capital of France?
Paris

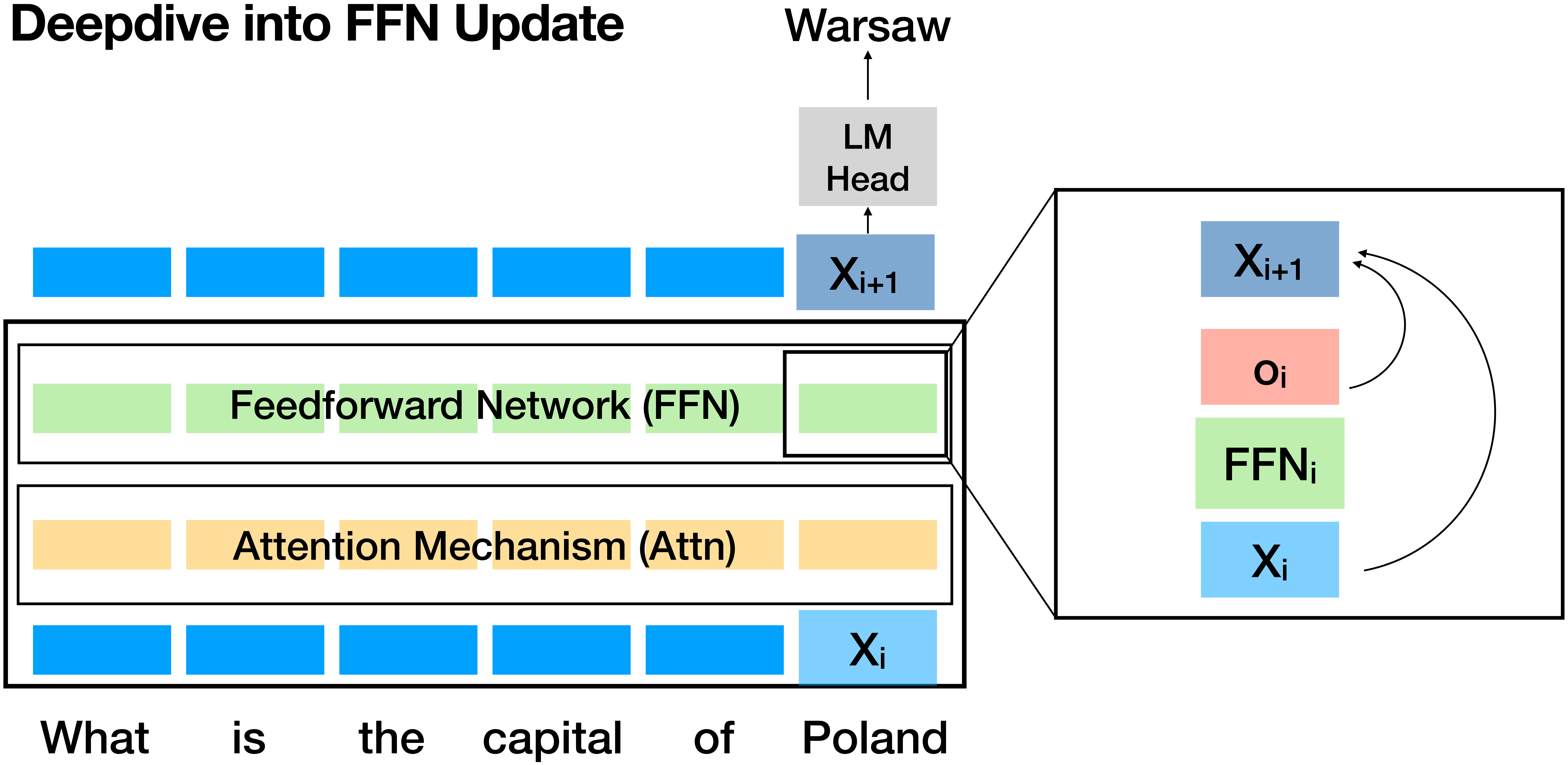
What is the capital of Poland?
Warsaw

`f | f(France) = Paris`

`f(Poland) = Warsaw`

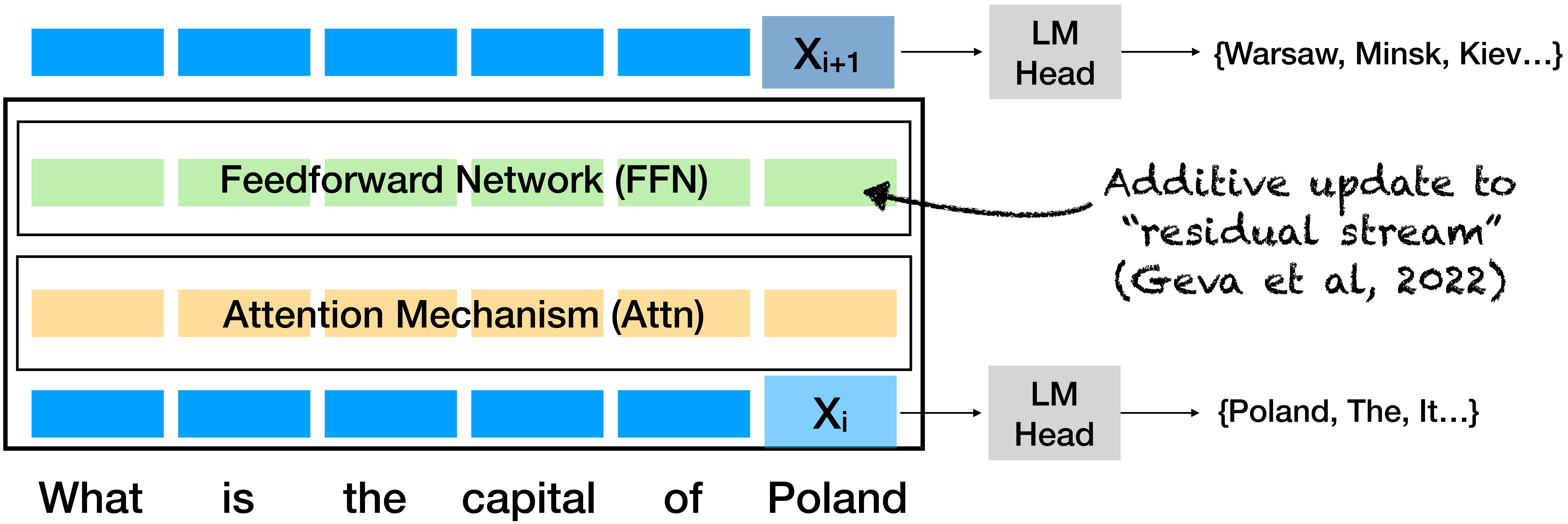
Abstract Functions in LLMs

Deepdive into FFN Update



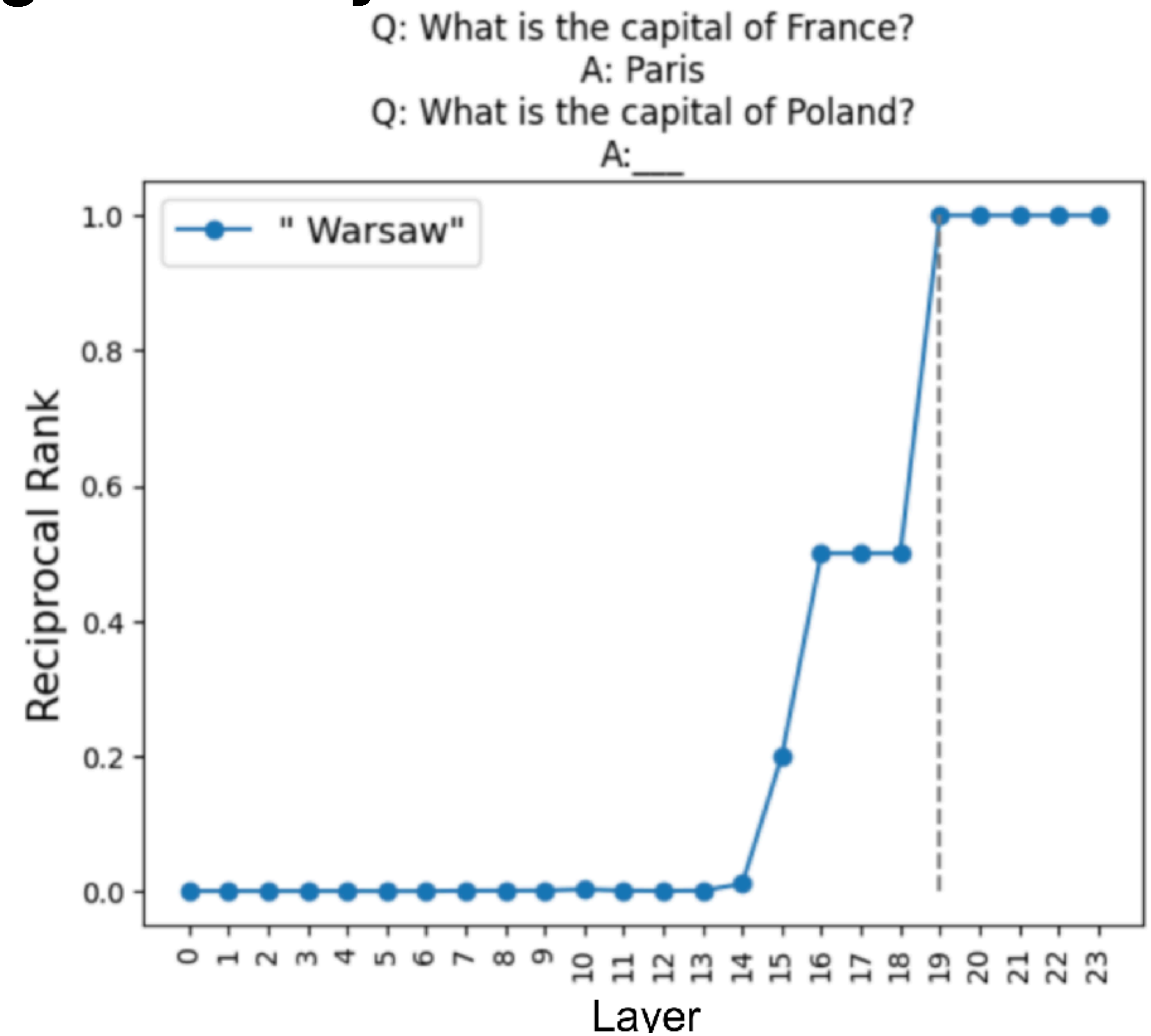
Abstract Functions in LLMs

Deepdive into FFN Update



Abstract Functions in LLMs

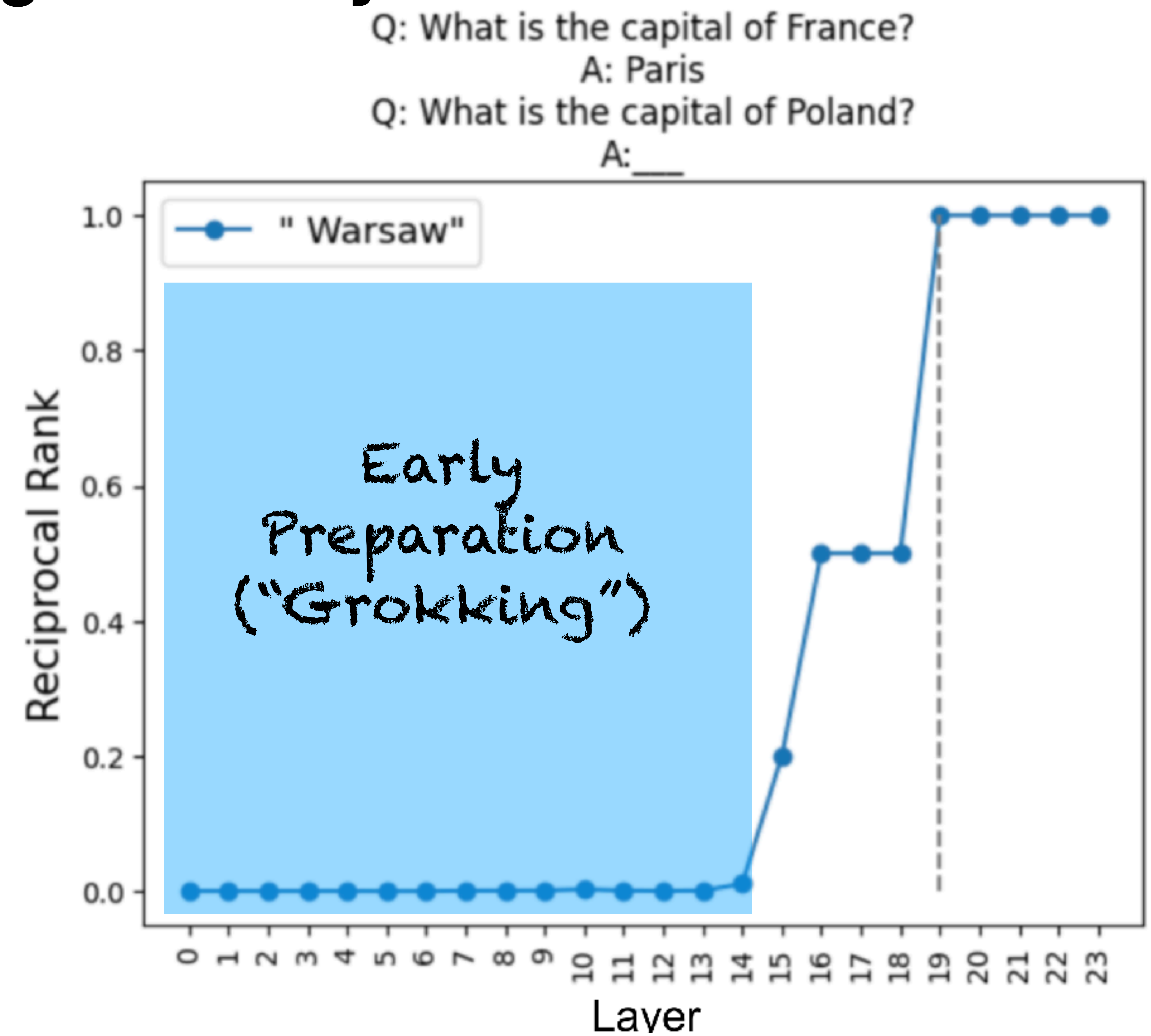
Processing across layers



Layer	Top Token
0	(
1	A
2	A
3	A
4	A
5	A
6	No
7	C
8	A
9	A
10	A
11	A
12	Unknown
13	C
14	St
15	Poland
16	Poland
17	Poland
18	Poland
19	Warsaw
20	Warsaw
21	Warsaw
22	Warsaw
23	Warsaw

Abstract Functions in LLMs

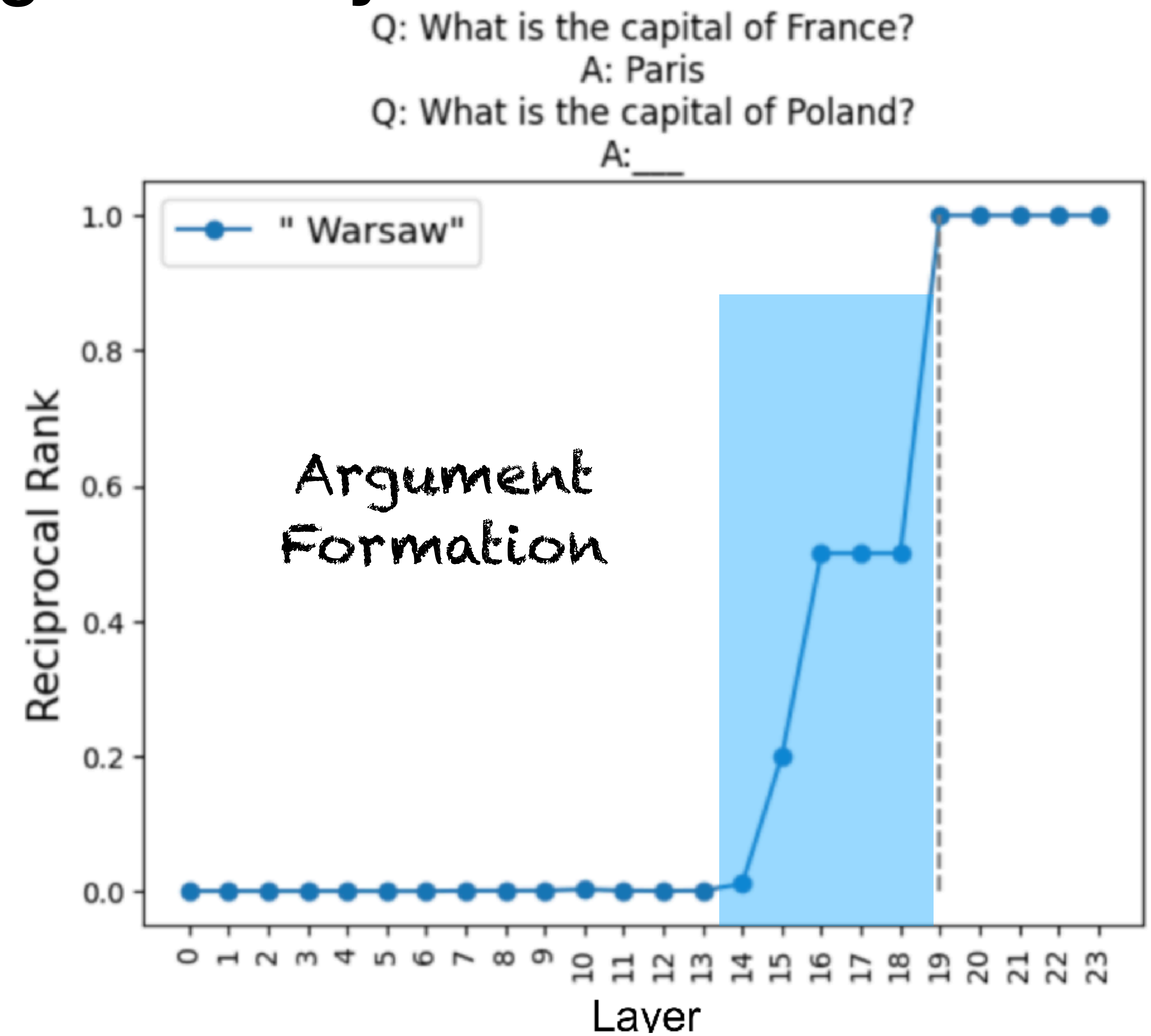
Processing across layers



Layer	Top Token
0	(
1	A
2	A
3	A
4	A
5	A
6	No
7	C
8	A
9	A
10	A
11	A
12	Unknown
13	C
14	St
15	Poland
16	Poland
17	Poland
18	Poland
19	Warsaw
20	Warsaw
21	Warsaw
22	Warsaw
23	Warsaw

Abstract Functions in LLMs

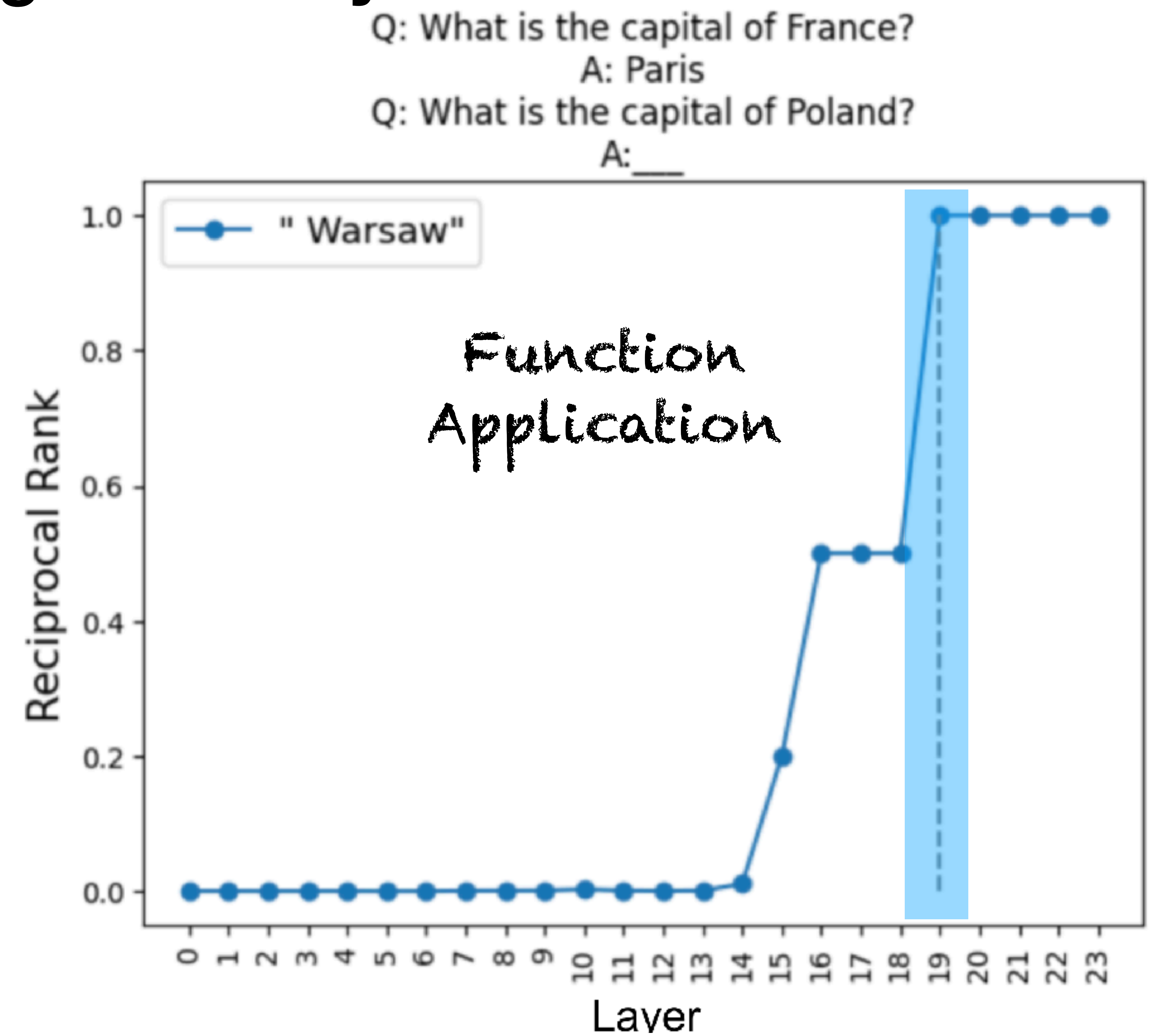
Processing across layers



Layer	Top Token
0	(
1	A
2	A
3	A
4	A
5	A
6	No
7	C
8	A
9	A
10	A
11	A
12	Unknown
13	C
14	St
15	Poland
16	Poland
17	Poland
18	Poland
19	Warsaw
20	Warsaw
21	Warsaw
22	Warsaw
23	Warsaw

Abstract Functions in LLMs

Processing across layers

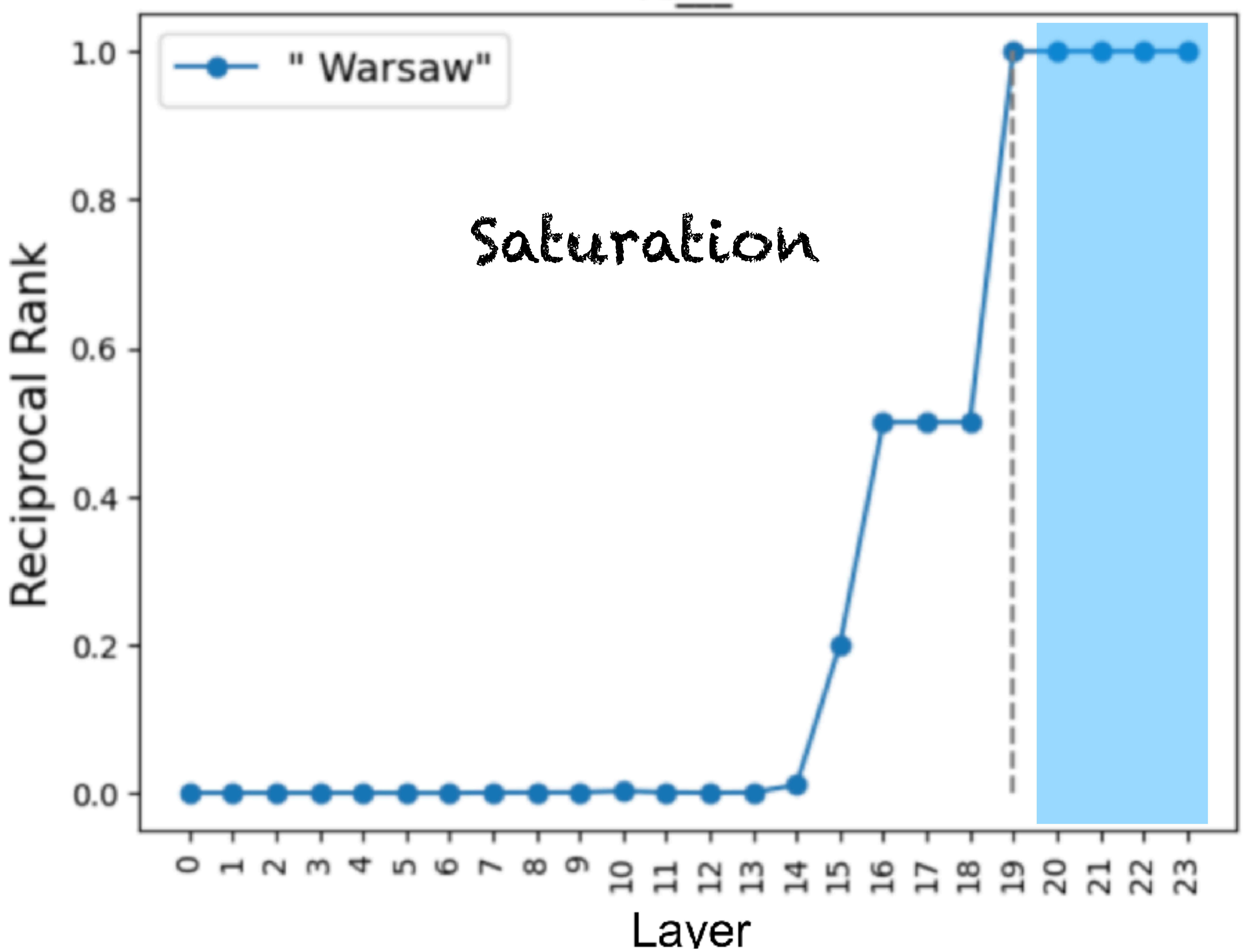


Layer	Top Token
0	(
1	A
2	A
3	A
4	A
5	A
6	No
7	C
8	A
9	A
10	A
11	A
12	Unknown
13	C
14	St
15	Poland
16	Poland
17	Poland
18	Poland
19	Warsaw
20	Warsaw
21	Warsaw
22	Warsaw
23	Warsaw

Abstract Functions in LLMs

Processing across layers

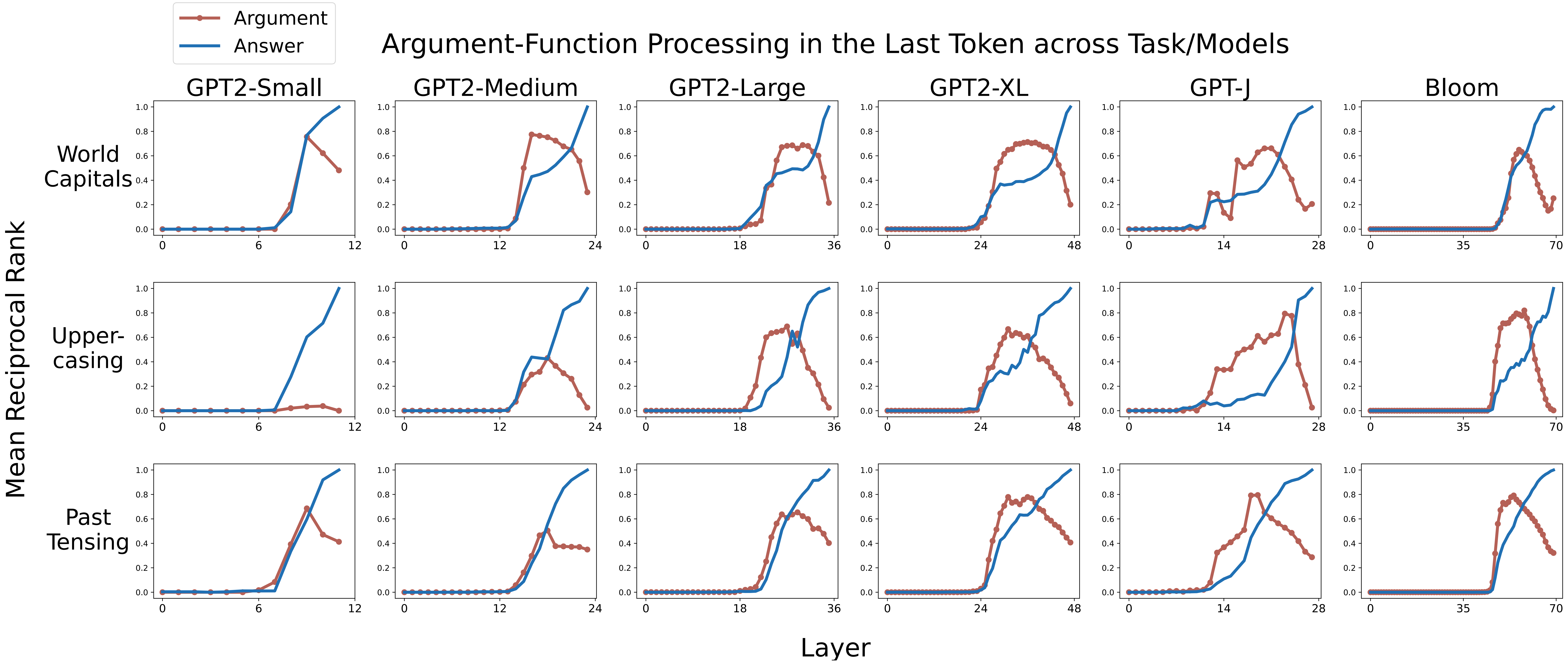
Q: What is the capital of France?
A: Paris
Q: What is the capital of Poland?
A:___



Layer	Top Token
0	(
1	A
2	A
3	A
4	A
5	A
6	No
7	C
8	A
9	A
10	A
11	A
12	Unknown
13	C
14	St
15	Poland
16	Poland
17	Poland
18	Poland
19	Warsaw
20	Warsaw
21	Warsaw
22	Warsaw
23	Warsaw

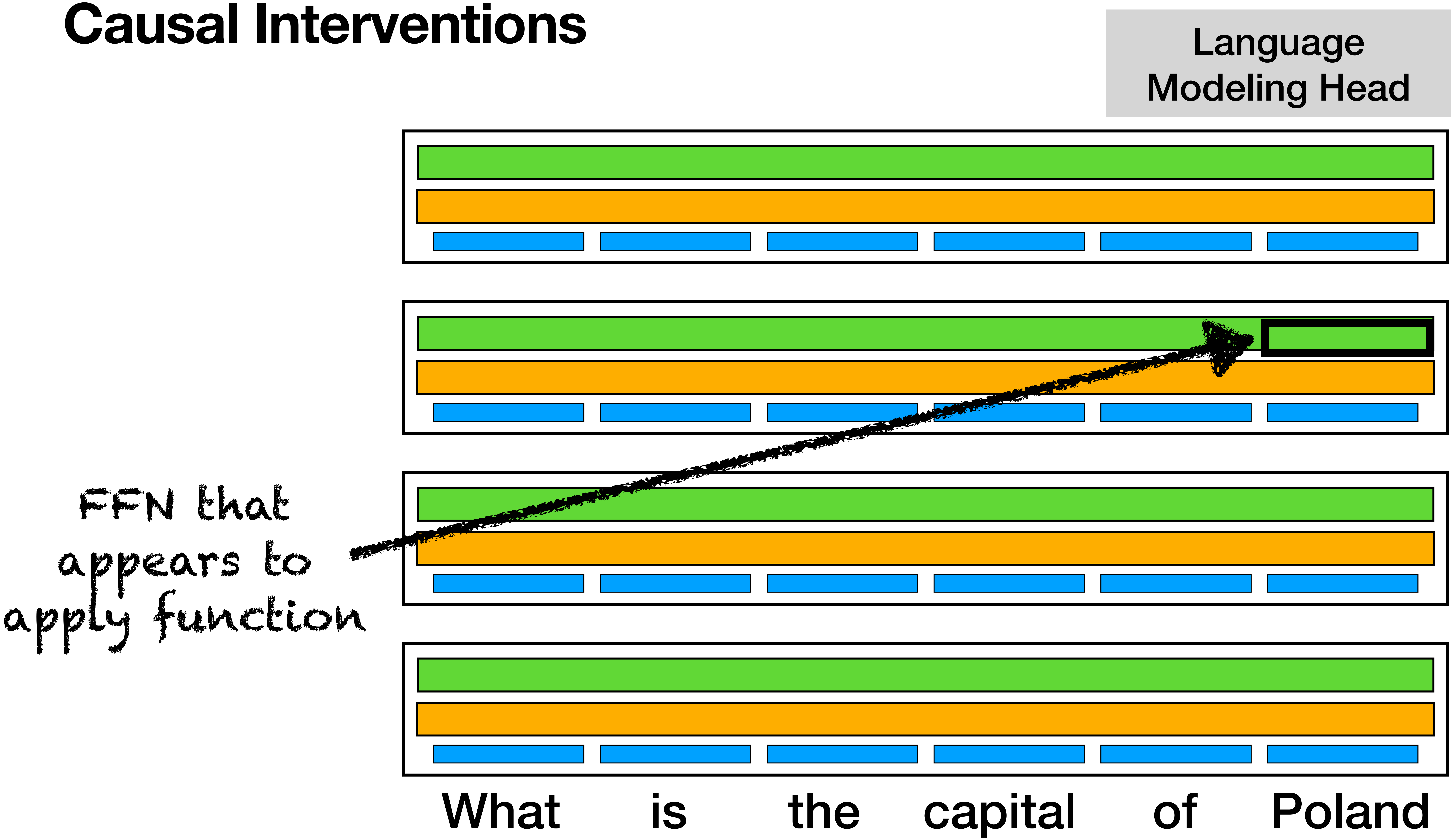
Abstract Functions in LLMs

Processing across layers



Abstract Functions in LLMs

Causal Interventions

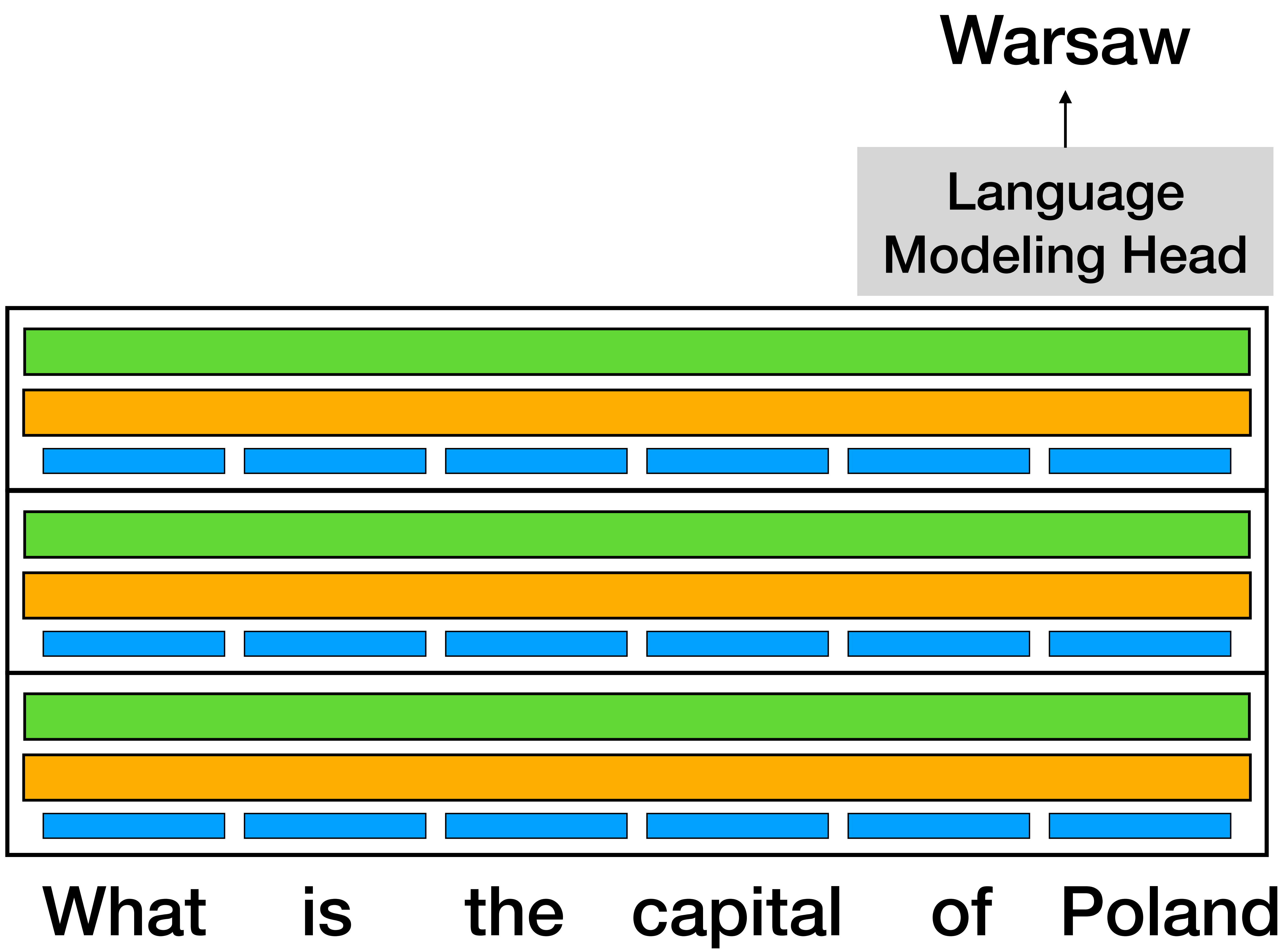


Abstract Functions in LLMs

Causal Interventions

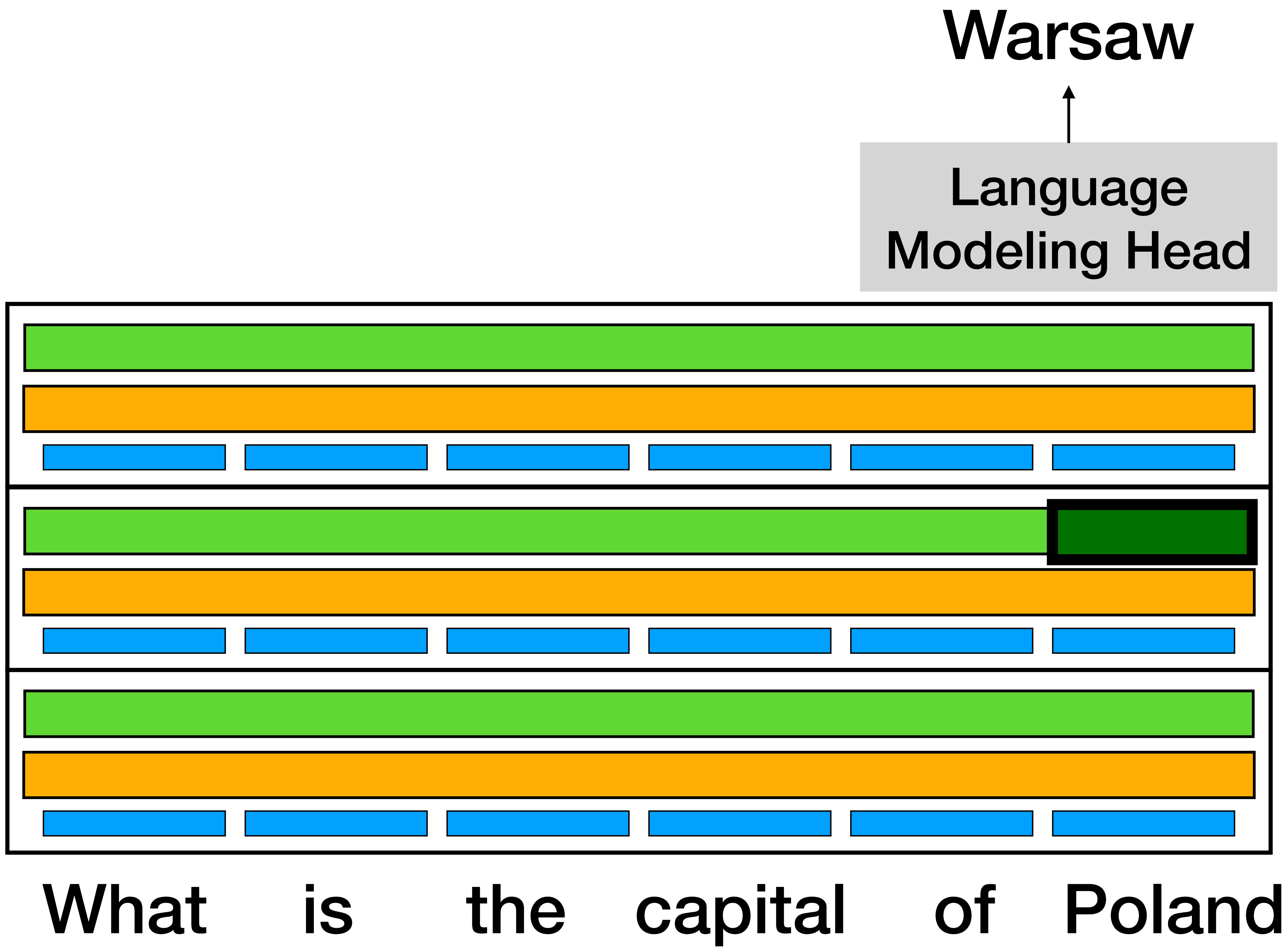
Abstract Functions in LLMs

Causal Interventions



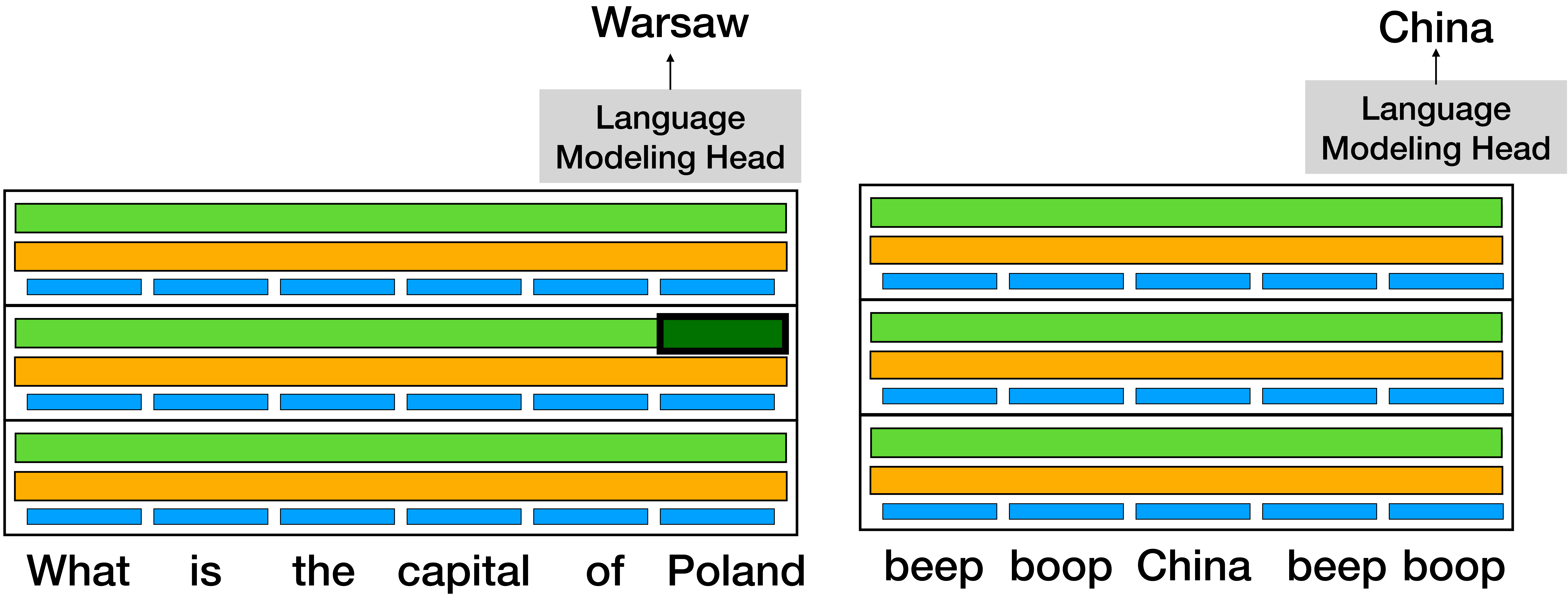
Abstract Functions in LLMs

Causal Interventions



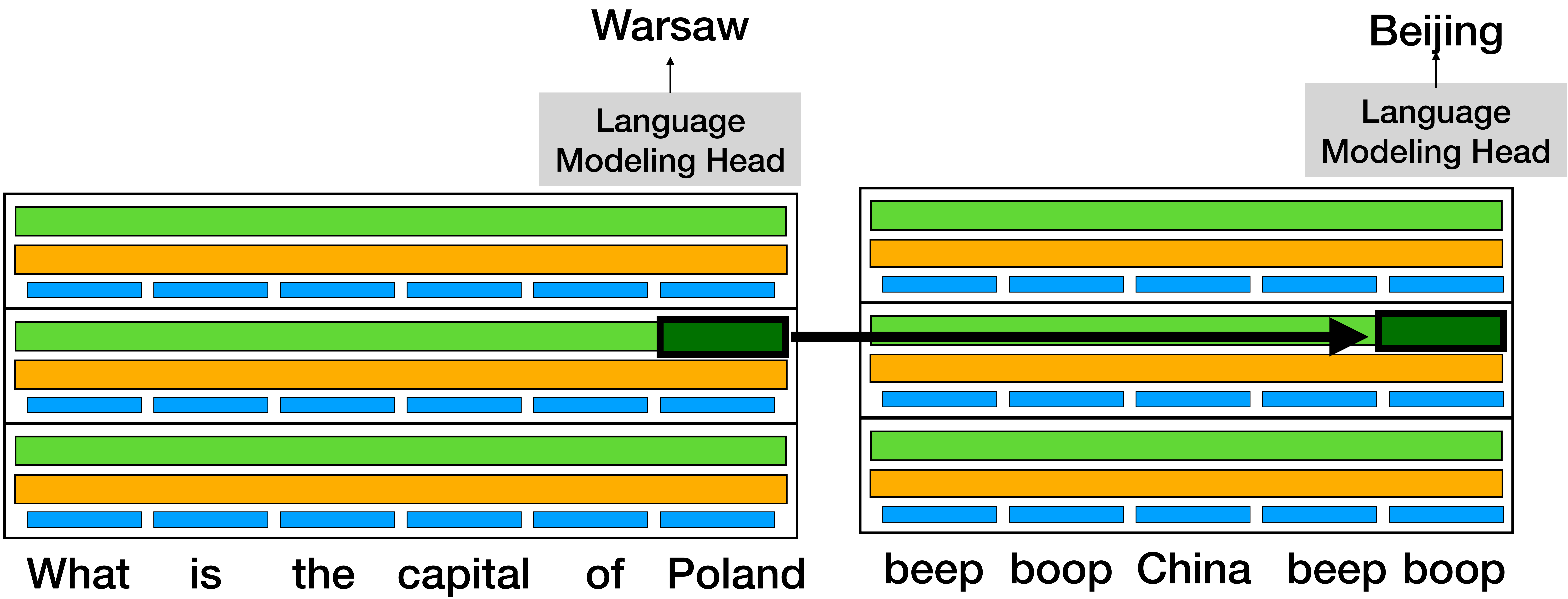
Abstract Functions in LLMs

Causal Interventions



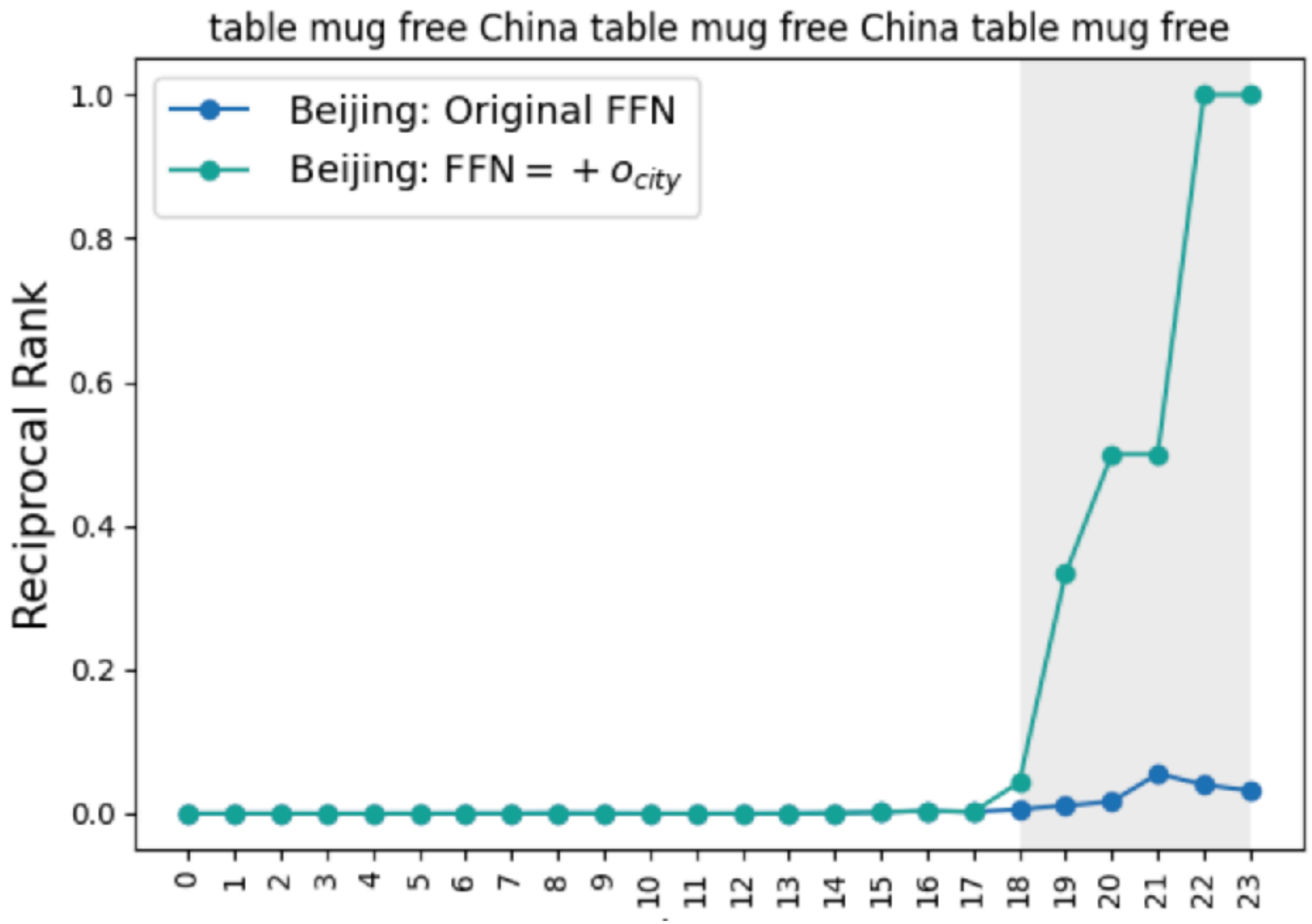
Abstract Functions in LLMs

Causal Interventions



Abstract Functions in LLMs

Causal Interventions



Consistently applies the same function, even for new arguments.

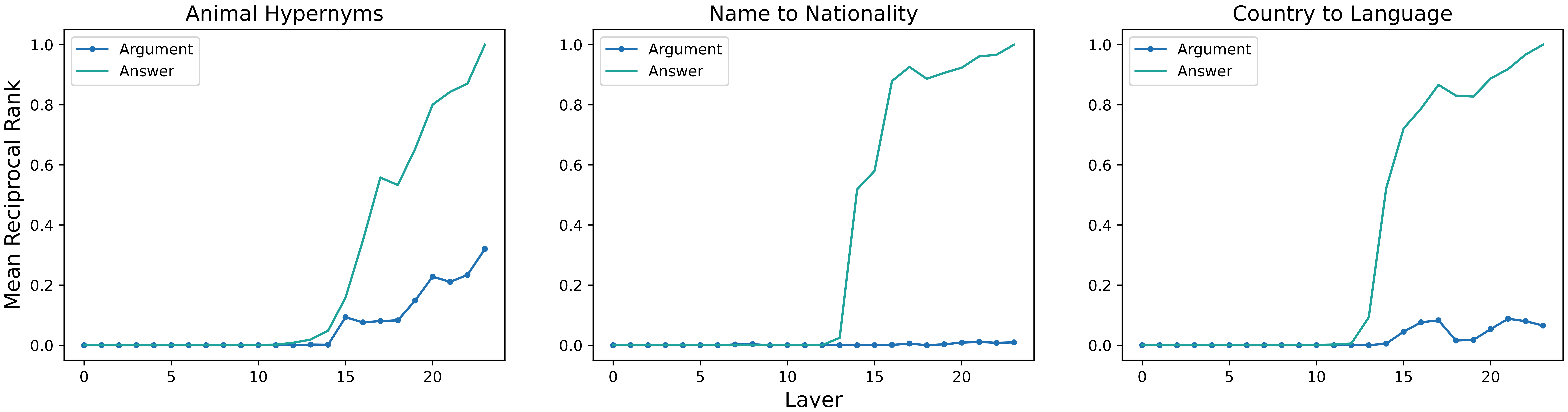
Abstract Functions in LLMs

Causal Interventions

Same pattern form many tasks (not just country->capital lookup)

Abstract Functions in LLMs

#not all relations



Though doesnt necessarily transfer to one-to-many or many-to-one relations

Abstract Functions in LLMs

#not all relations

Extractive

The capital of China is Beijing.
What is the capital of China?

Abstractive

What is the capital of China?

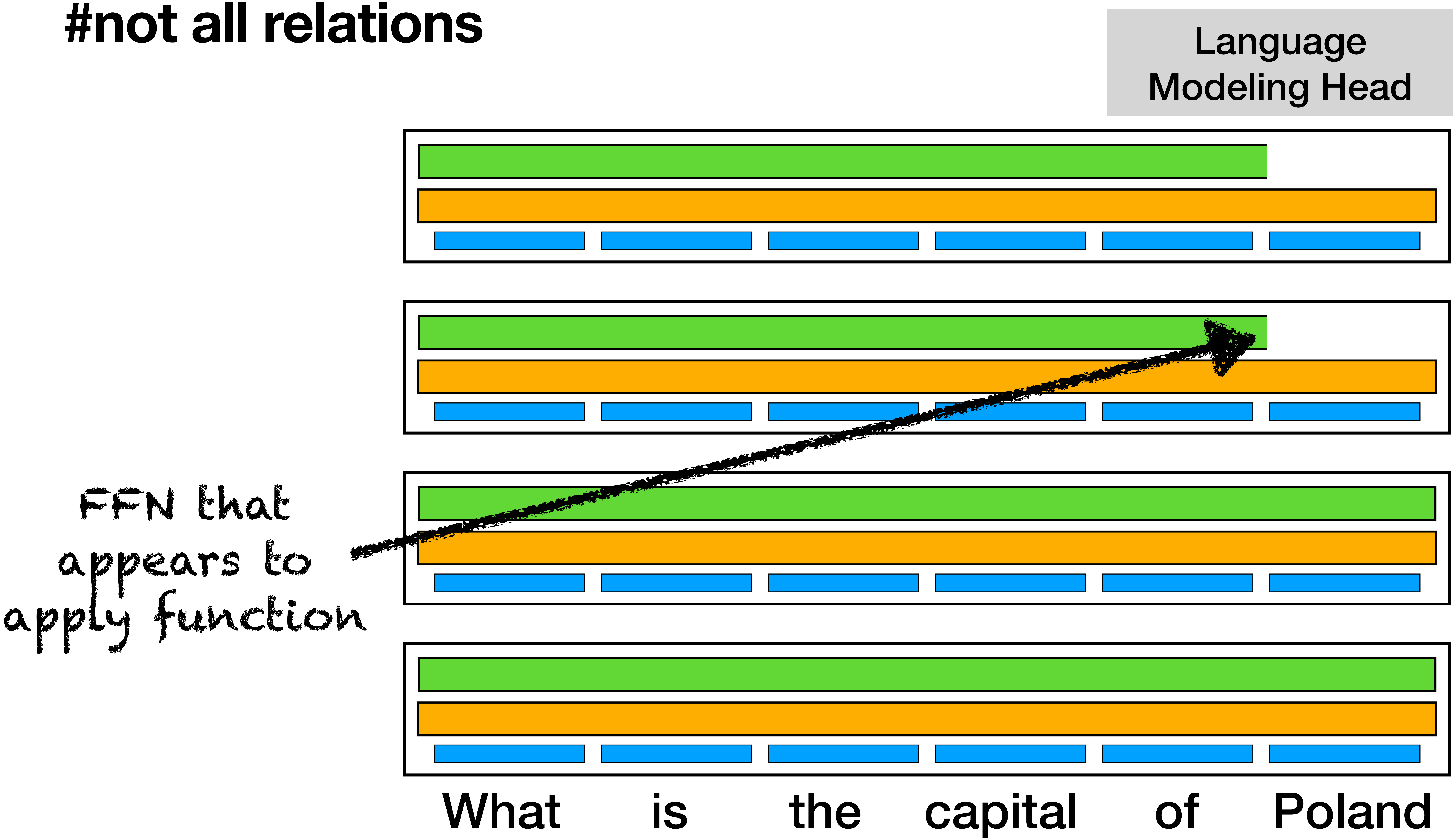
Abstract Functions in LLMs

#not all relations

Extractive	Abstractive
The capital of China is **Beijing**. What is the capital of China?	What is the capital of China?

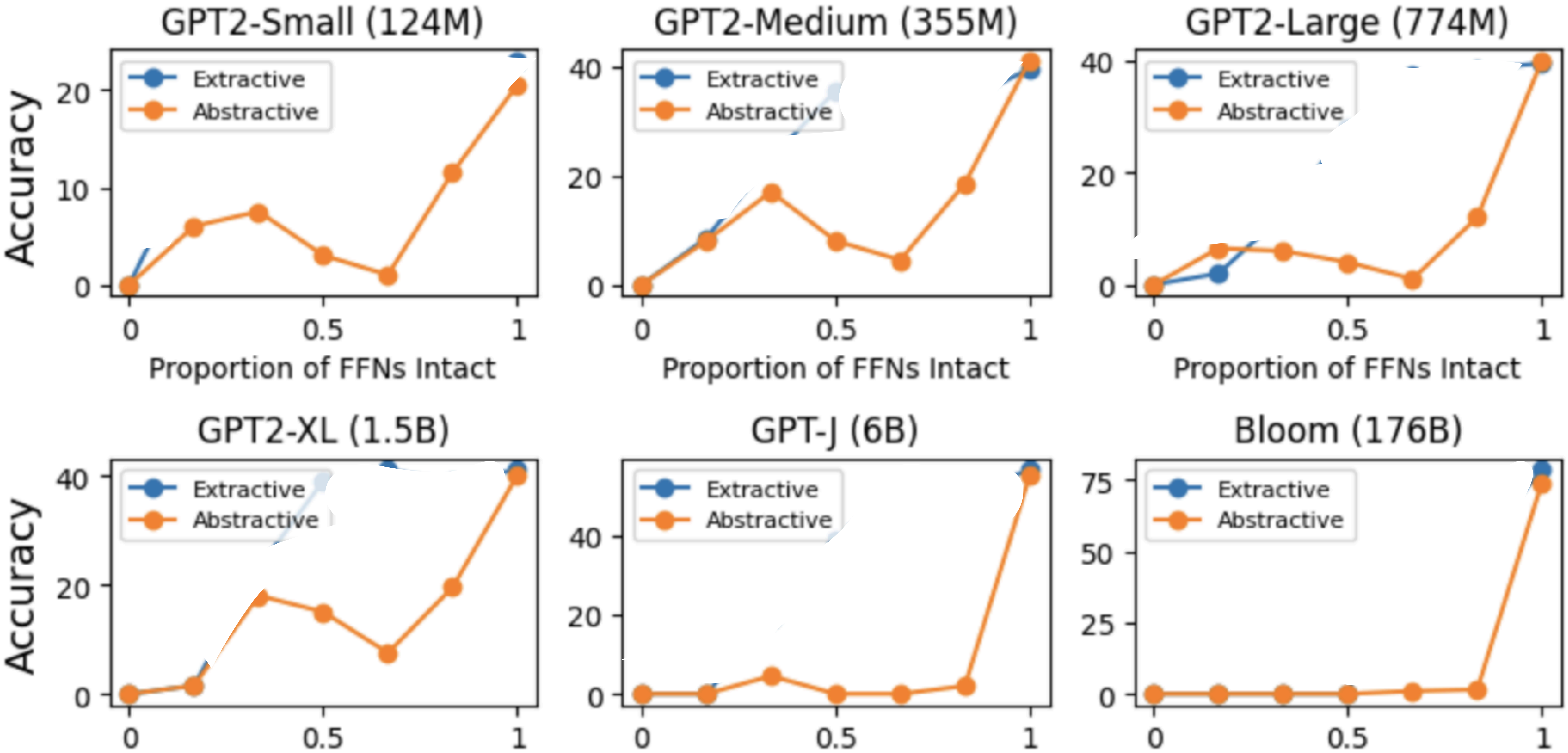
Abstract Functions in LLMs

#not all relations



Abstract Functions in LLMs

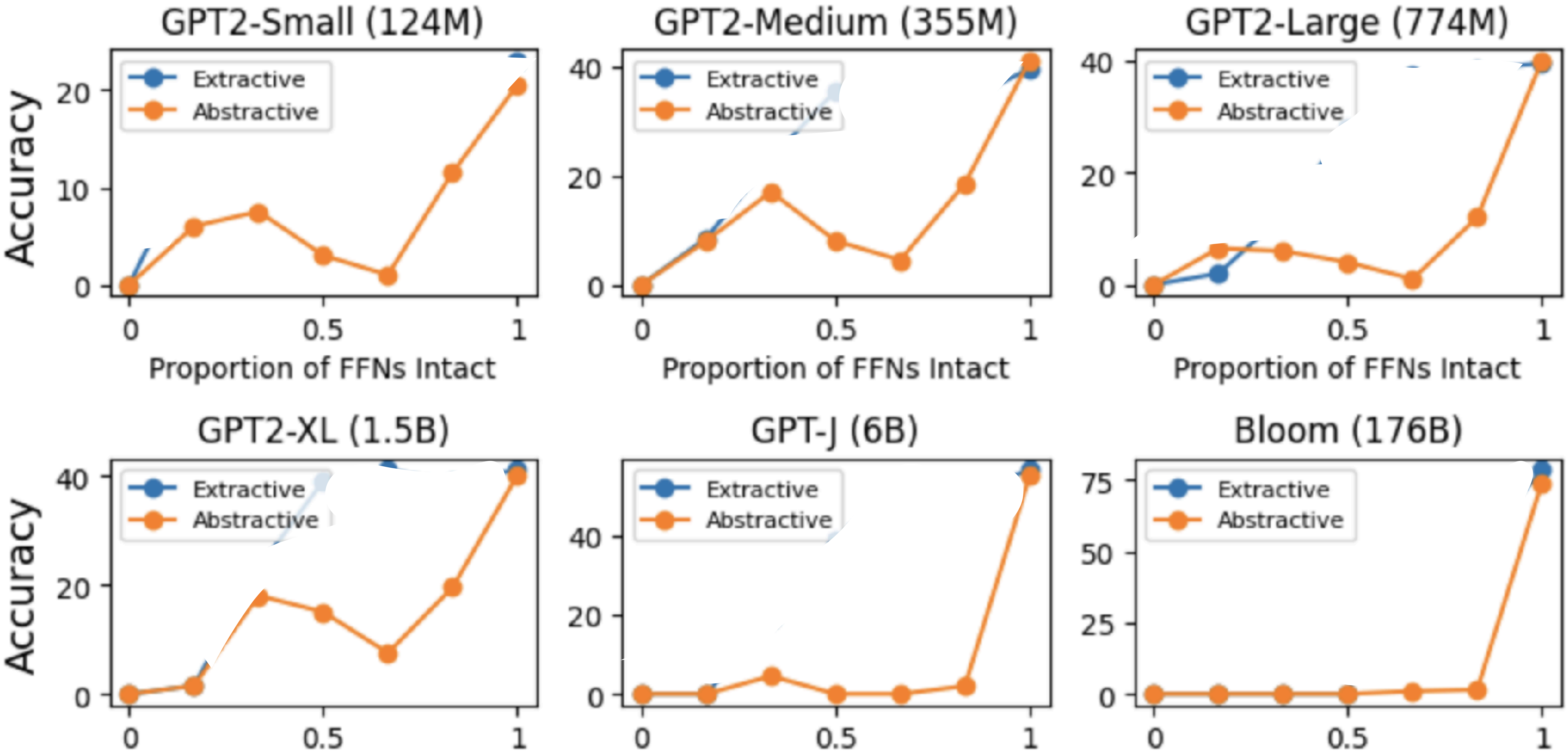
#not all relations



Abstract Functions in LLMs

#not all relations

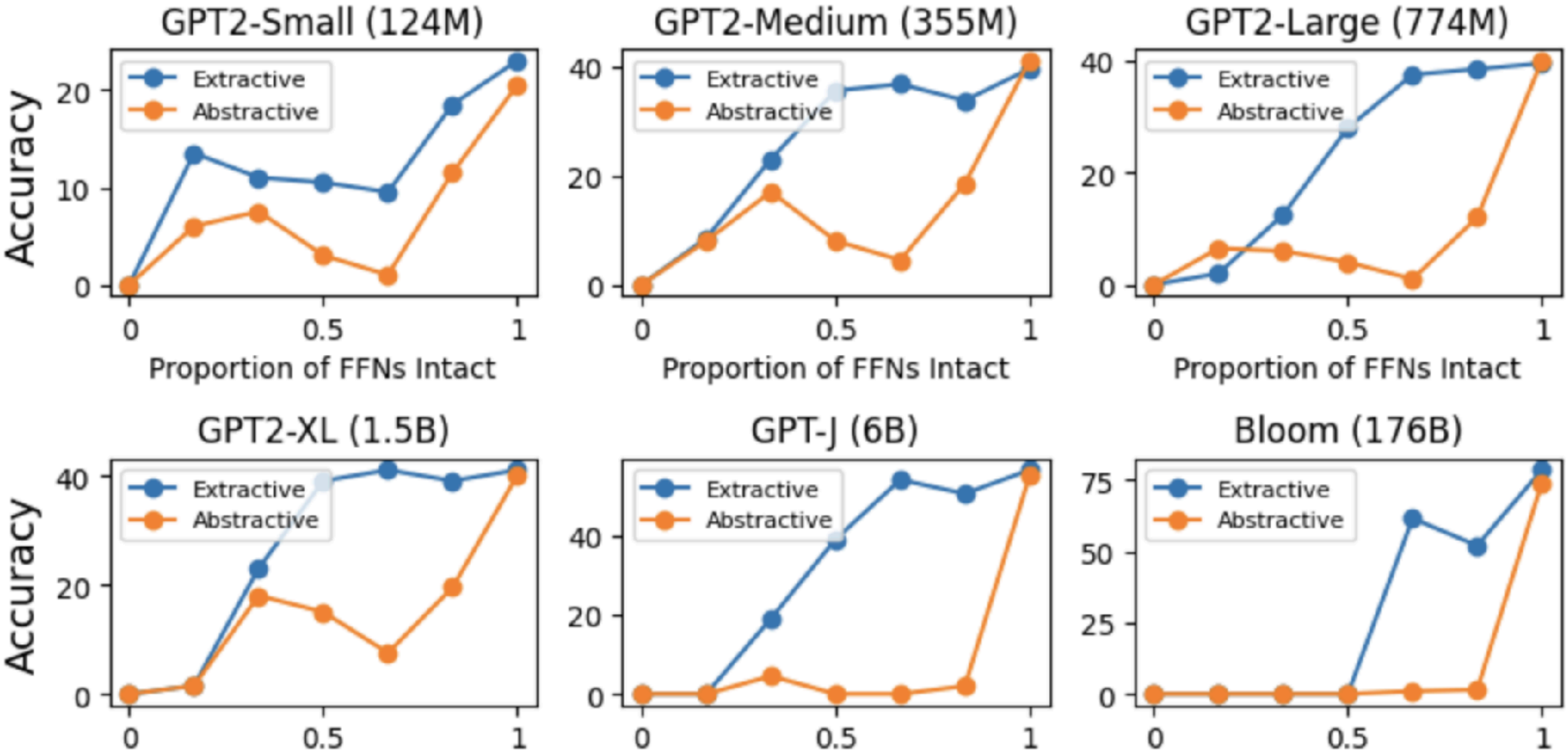
FFNs necessary for performance on Abstractive task



Abstract Functions in LLMs

#not all relations

But play no role in Extractive task



Abstract Functions in LLMs

Different Mechanism in Abstractive vs. Extractive Settings

Extractive

The capital of China is **Warsaw**.
What is the capital of China?

Abstractive

What is the capital of China?

Abstract Functions in LLMs

Different Mechanism in Abstractive vs. Extractive Settings

Extractive

The capital of China is **Warsaw**.
What is the capital of China?

Abstractive

What is the capital of China?

could compete with each other!

Abstract Functions in LLMs

Different Mechanism in Abstractive vs. Extractive Settings

E

Characterizing Mechanisms for Factual Recall in Language Models

Qinan Yu Jack Merullo Ellie Pavlick
Brown University
Department of Computer Science
{qinan_yu,jack_merullo,ellie_pavlick}@brown.edu

The capital

What is the capital of China?

f China?

could compete with each other!

Abstract Functions in LLMs

Different Mechanism in Abstractive vs. Extractive Settings

The capital of Poland is London.
What is the capital of Poland?

Abstract Functions in LLMs

Different Mechanism in Abstractive vs. Extractive Settings

Country

The capital of Poland is London.
What is the capital of Poland?

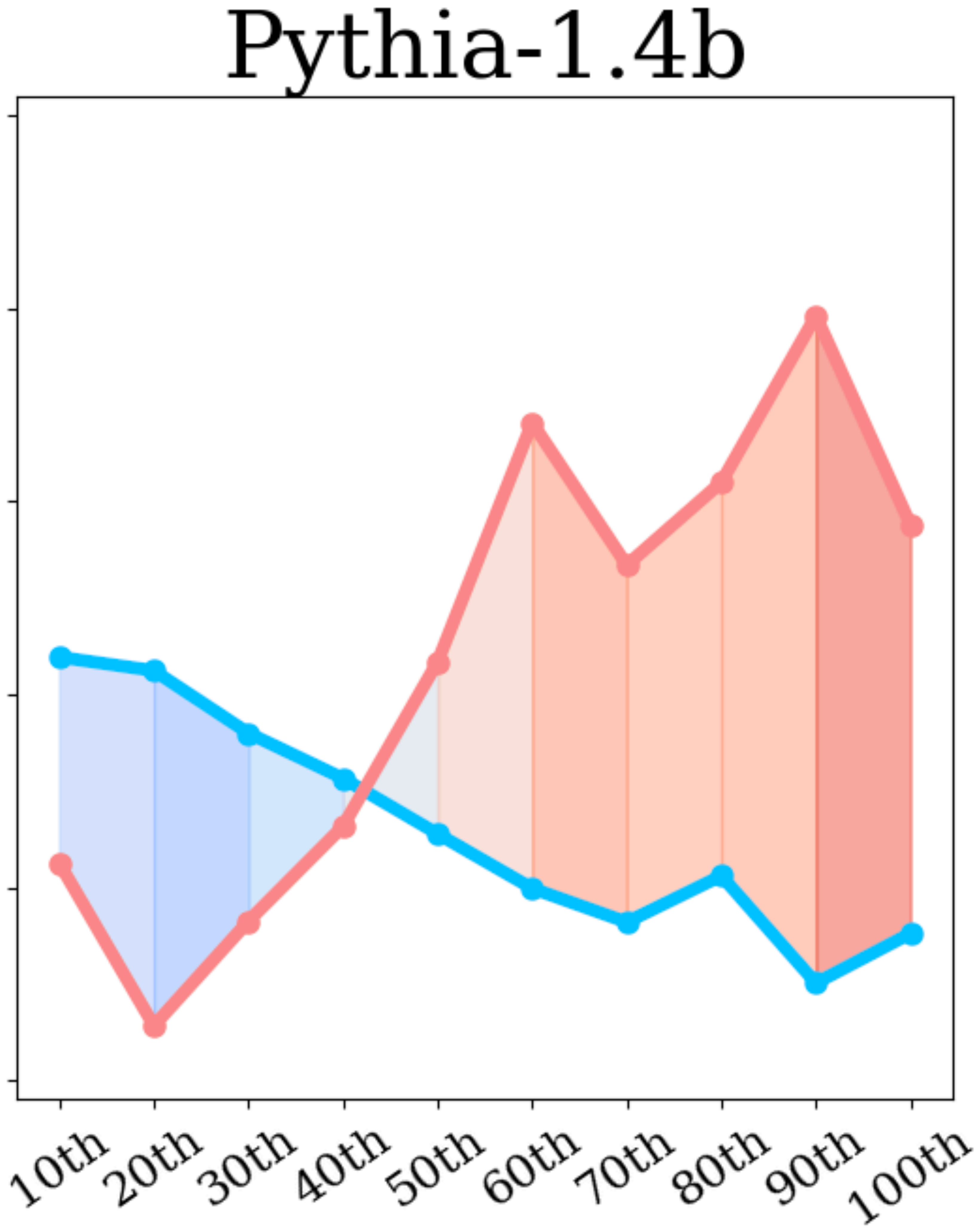
London In-Context Answer

Warsaw Memorized Answer

Abstract Functions in LLMs

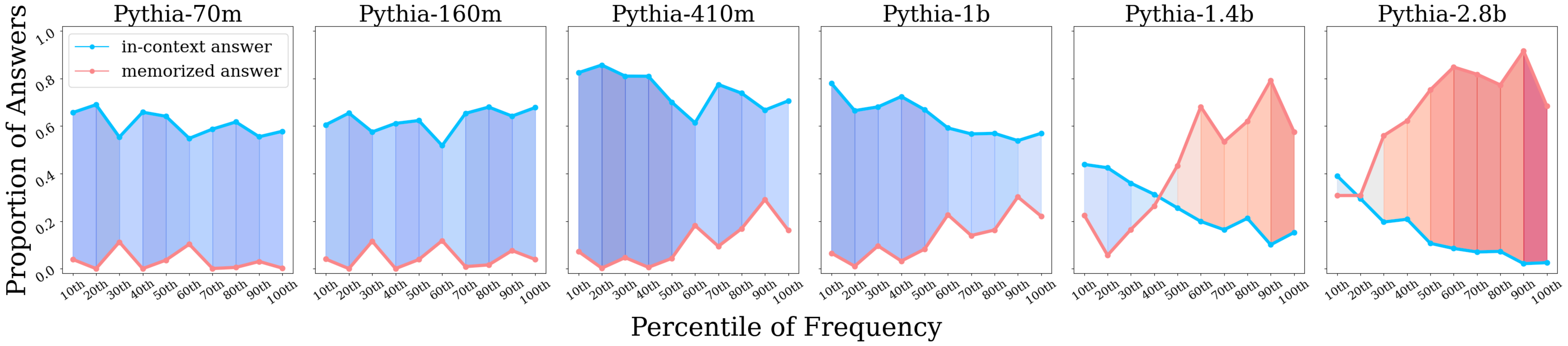
Training data frequency affects which mechanism is used

As the count of country increases, model is more likely to predict the memorized answer and less likely to predict the in-context one



Abstract Functions in LLMs

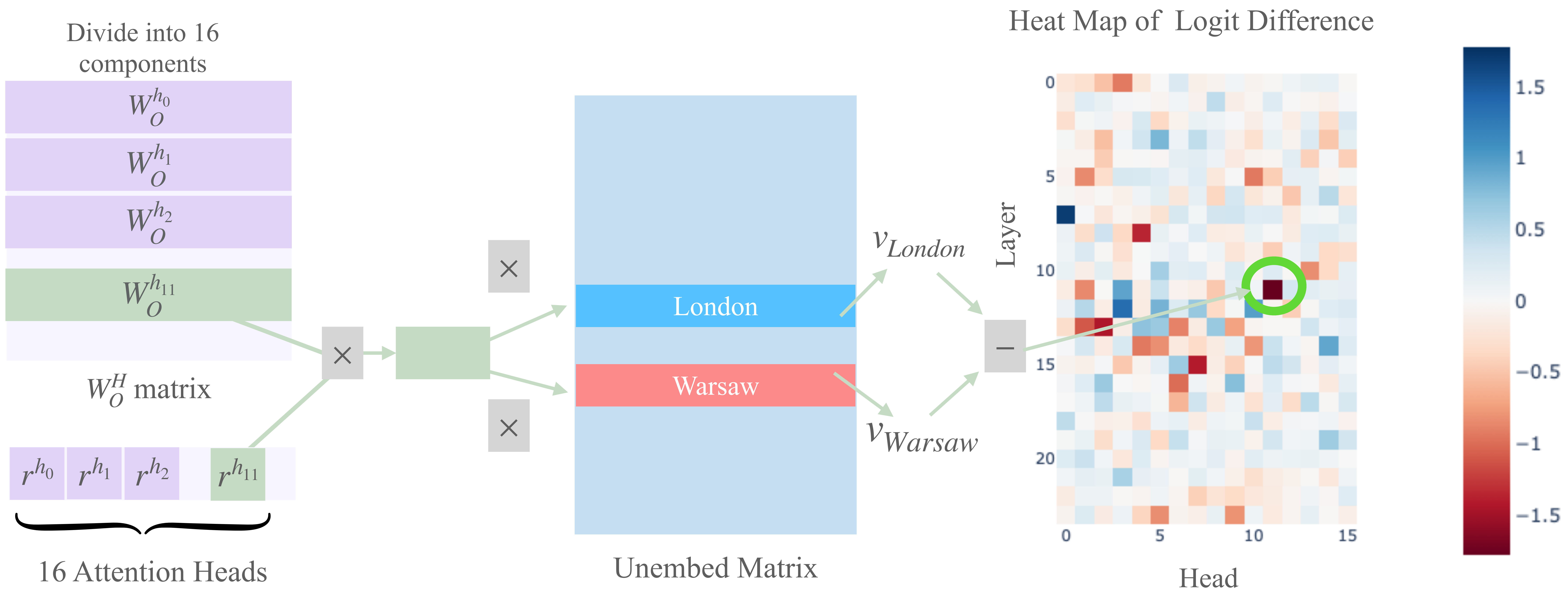
Training data frequency affects which mechanism is used



Trend appears to be associated with model size. Larger models prefer memorized answers, but change affects frequent countries first...

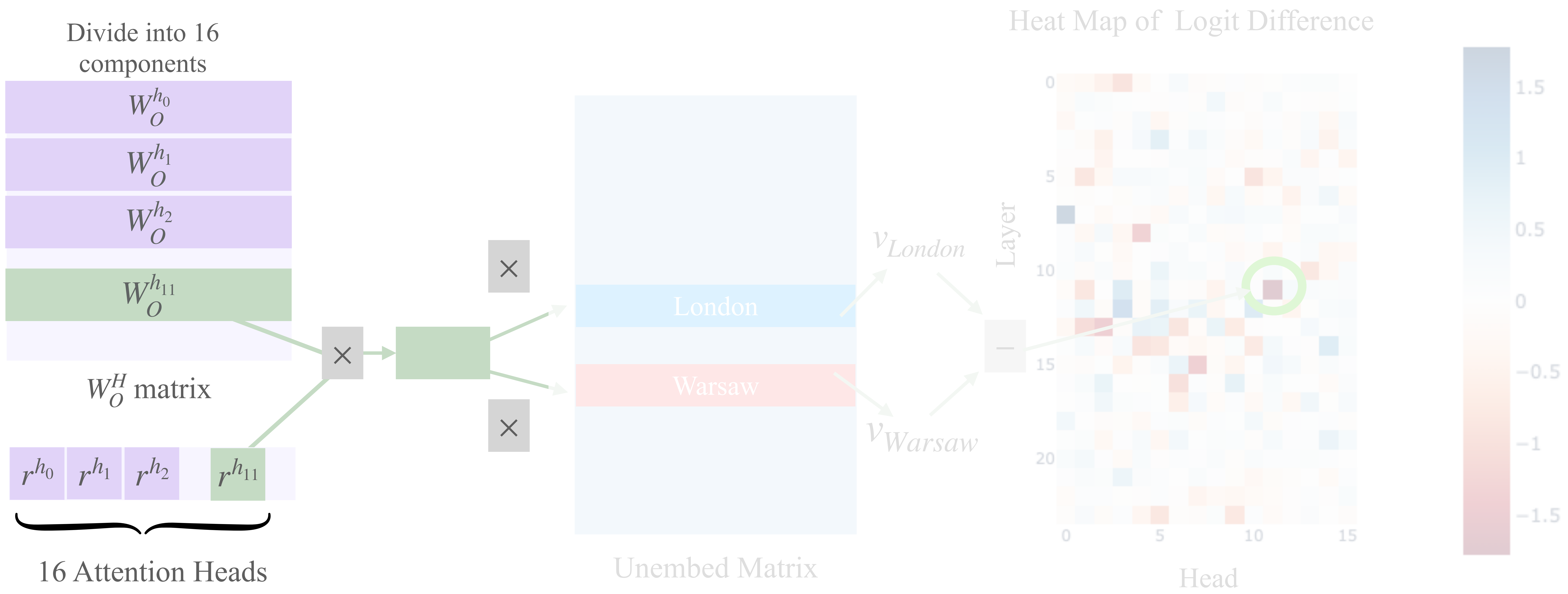
Abstract Functions in LLMs

Path Patching to Locate Important Attention Heads (Wang et al.)



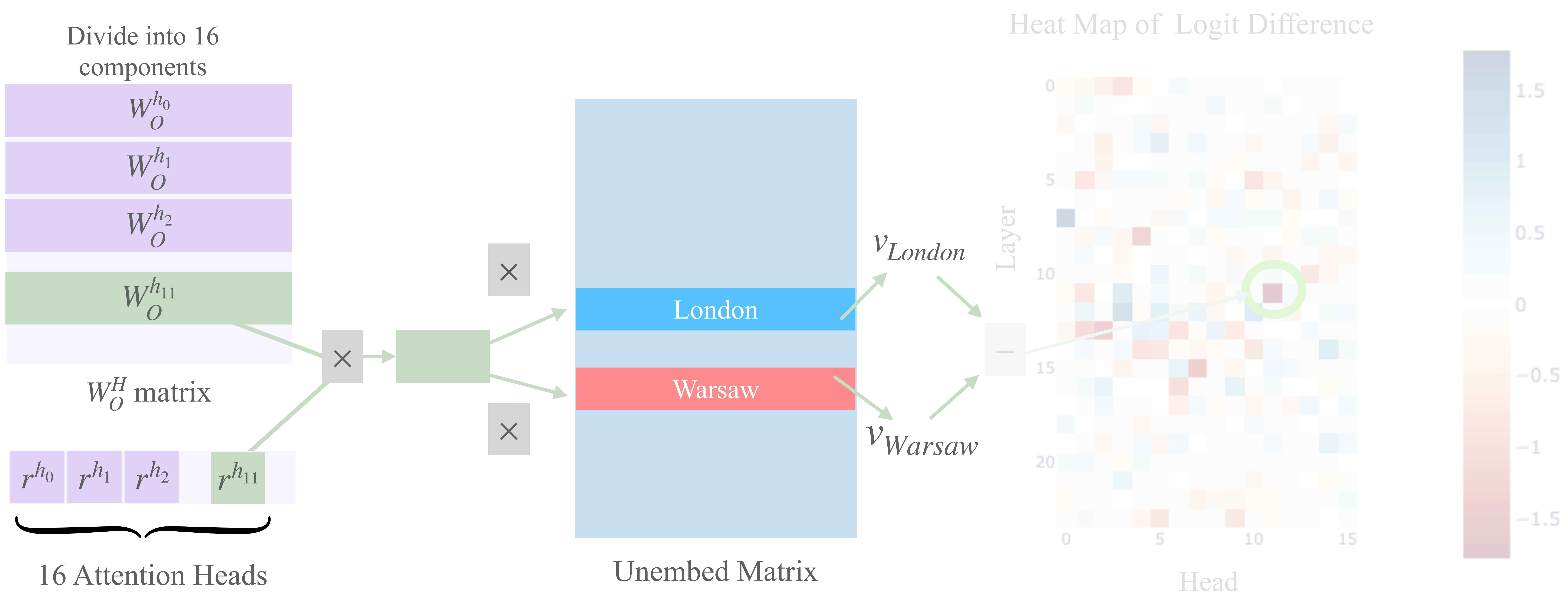
Abstract Functions in LLMs

Path Patching to Locate Important Attention Heads (Wang et al.)



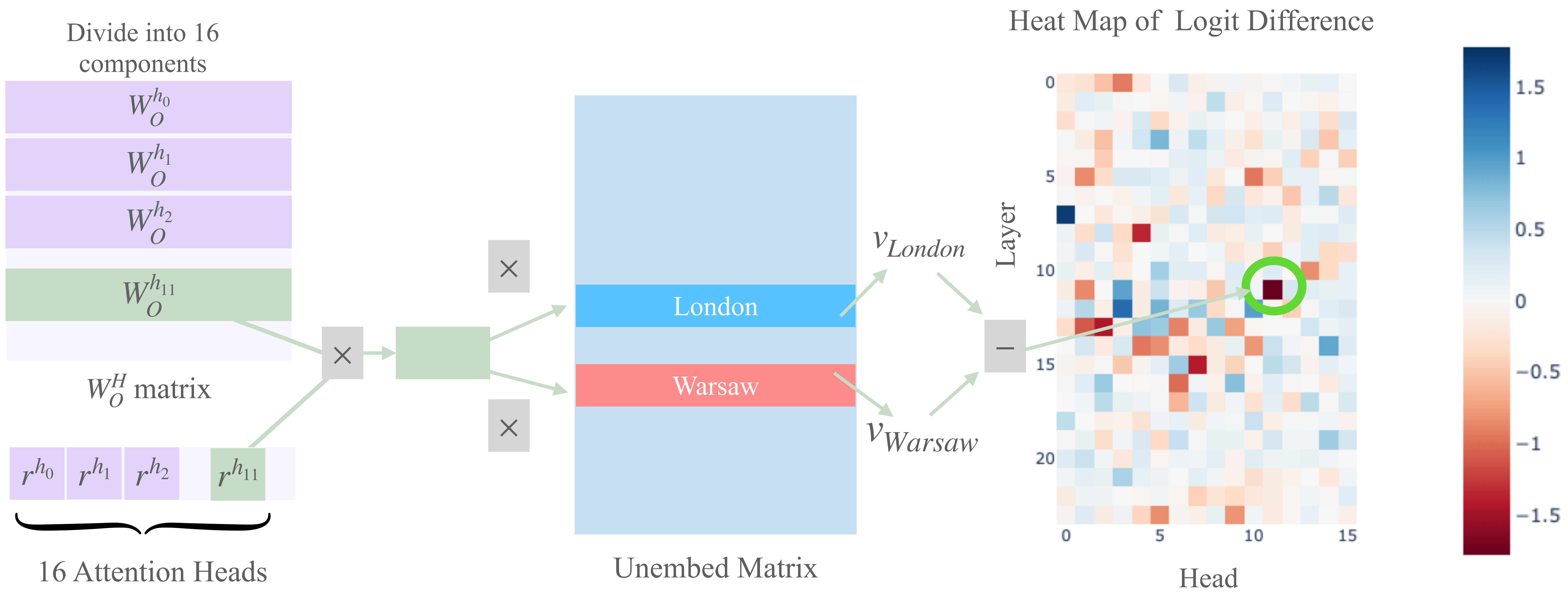
Abstract Functions in LLMs

Path Patching to Locate Important Attention Heads (Wang et al.)



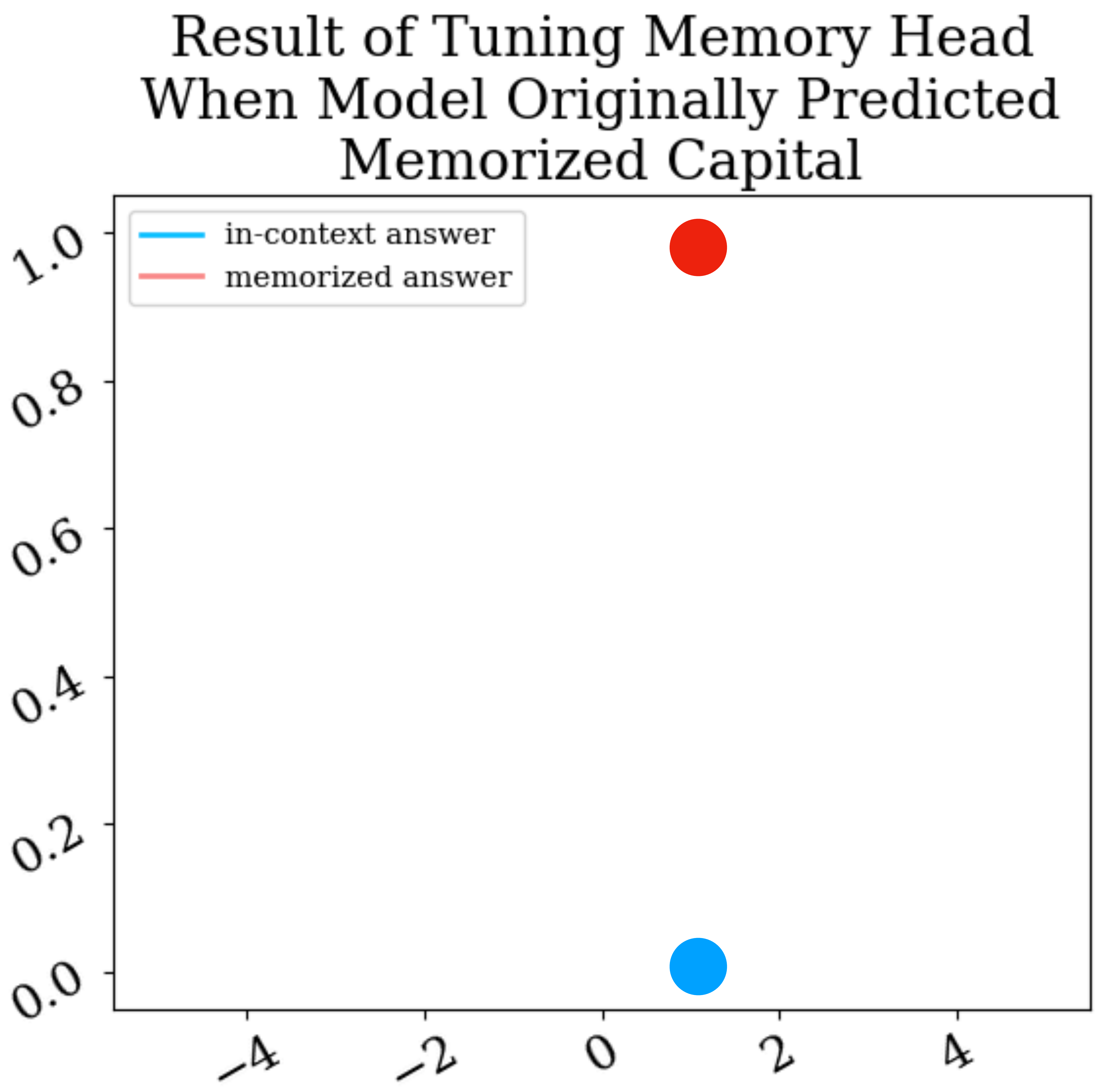
Abstract Functions in LLMs

Path Patching to Locate Important Attention Heads (Wang et al.)



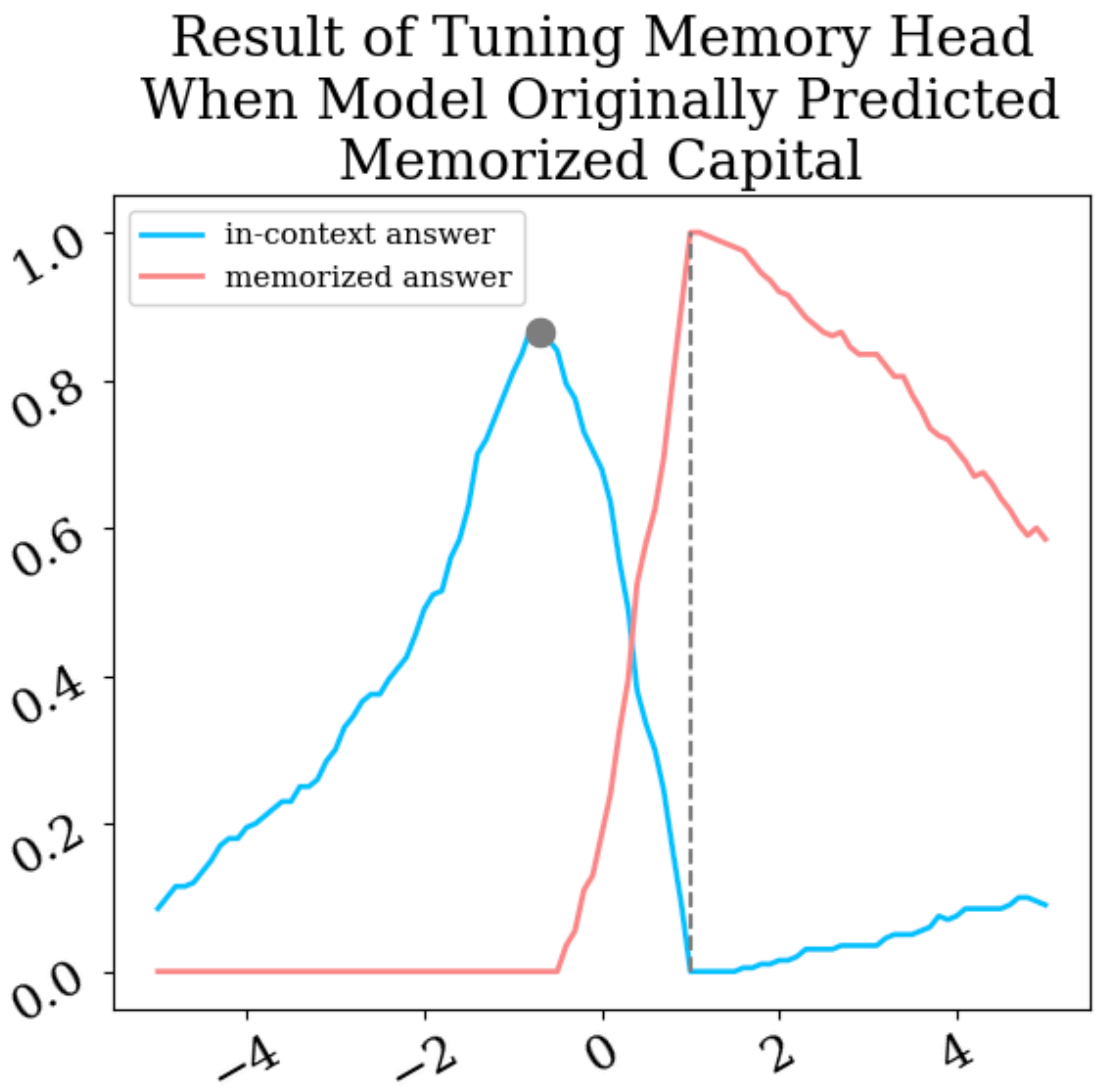
Abstract Functions in LLMs

Specific heads mediate which mechanism is used



Abstract Functions in LLMs

Specific heads mediate which mechanism is used



Abstract Functions in LLMs

Specific heads mediate which mechanism is used

Abstract Functions in LLMs

Summary and Discussion

- We focus on a simple but important step of language processing: retrieving factual information from memory
- We find that Transformer LLMs appear to implement this step using a simple linear update mechanism computer in the FFNs
- The computation is modular and generic. It can be transferred to new contexts in a zero-shot manner.
- It’s use is modulated by independent, local, and (somewhat) controllable mechanisms
- Serves as a proof of concept for how “black box” LLM behaviors can be translated into interpretable, functional terms

This Talk

- Transformers and the “Mental Model of LLMs”
- **Two Proofs of Concept:**
  - Abstract representation of relations
  - **Modular and reusable algorithmic “building blocks”**

Understanding LLM Circuits and Algorithms

Jack Merullo

Carsten Eickhoff

Published as a conference paper at ICLR 2024

CIRCUIT COMPONENT REUSE ACROSS TASKS IN TRANSFORMER LANGUAGE MODELS

Jack Merullo
Department of Computer Science
Brown University
jack_merullo@brown.edu

Carsten Eickhoff
School of Medicine
University of Tübingen
carsten.eickhoff@uni-tuebingen.de

Ellie Pavlick
Department of Computer Science
Brown University
ellie_pavlick@brown.edu

Talking Heads: Understanding Inter-layer Communication in Transformer Language Models

Anonymous Author(s)
Affiliation
Address
email

Understanding LLM Circuits and Algorithms

Two Different Language Processing Tasks

Then, Matthew and Robert had a lot of fun at the school.
Robert gave a ring to _____

Wang et al. (2022)

Q: One the table, there is a blue pencil, a black necklace, and a yellow lighter.
What color is the pencil?
A: _____

Ippolito and Callison-Burch (2023)

Understanding LLM Circuits and Algorithms

Prior Work: The IOI Circuit

Matthew and Robert had a lot of fun at the school.
Robert gave a ring to _____.

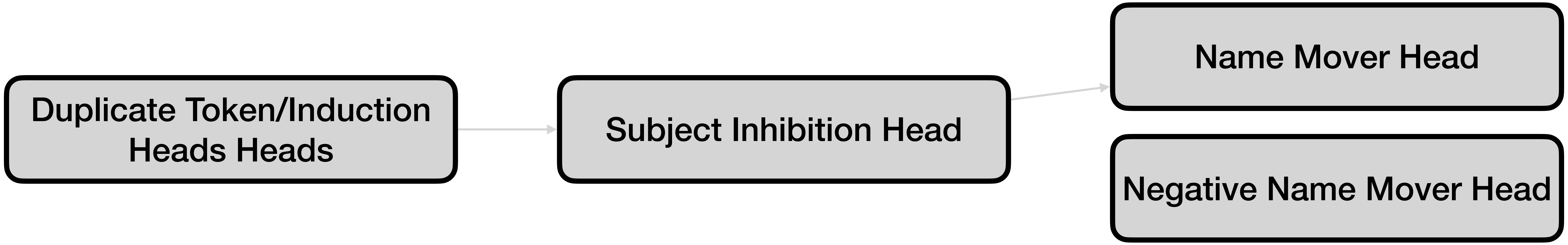
INTERPRETABILITY IN THE WILD: A CIRCUIT FOR INDIRECT OBJECT IDENTIFICATION IN GPT-2 SMALL

Kevin Wang[1], **Alexandre Variengien**[1], **Arthur Conmy**[1], **Buck Shlegeris**[1] **& Jacob Steinhardt**[1,2]
[1]Redwood Research
[2]UC Berkeley
kevin@rdwrs.com, alexandre@rdwrs.com,
arthur@rdwrs.com, buck@rdwrs.com, jsteinhardt@berkeley.edu

Understanding LLM Circuits and Algorithms

Prior Work: The IOI Circuit

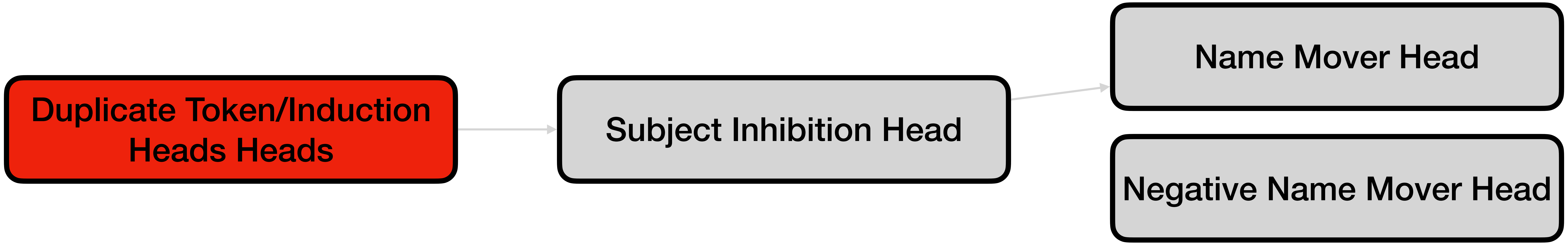
Matthew and Robert had a lot of fun at the school.
Robert gave a ring to _____.



Understanding LLM Circuits and Algorithms

Prior Work: The IOI Circuit

Matthew and Robert had a lot of fun at the school.
Robert gave a ring to _____.

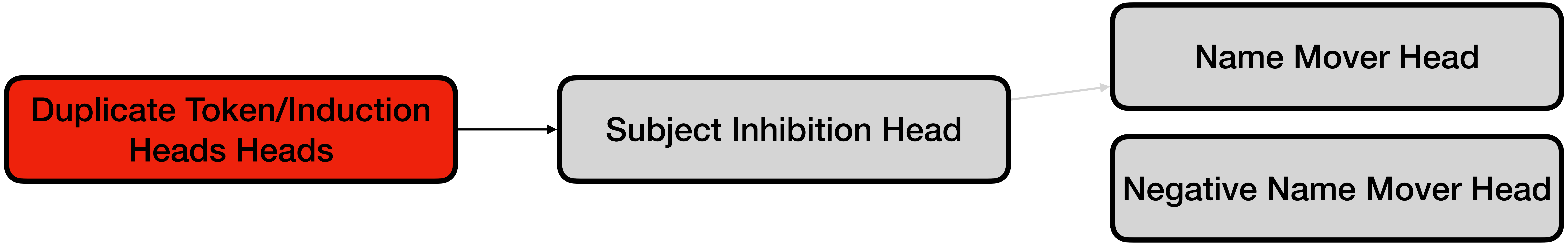


1. Identify any duplicated names.

Understanding LLM Circuits and Algorithms

Prior Work: The IOI Circuit

Matthew and Robert had a lot of fun at the school.
Robert gave a ring to _____.



1. Identify any duplicated names.
2. Alert the S-Inhibition head of their location

Understanding LLM Circuits and Algorithms

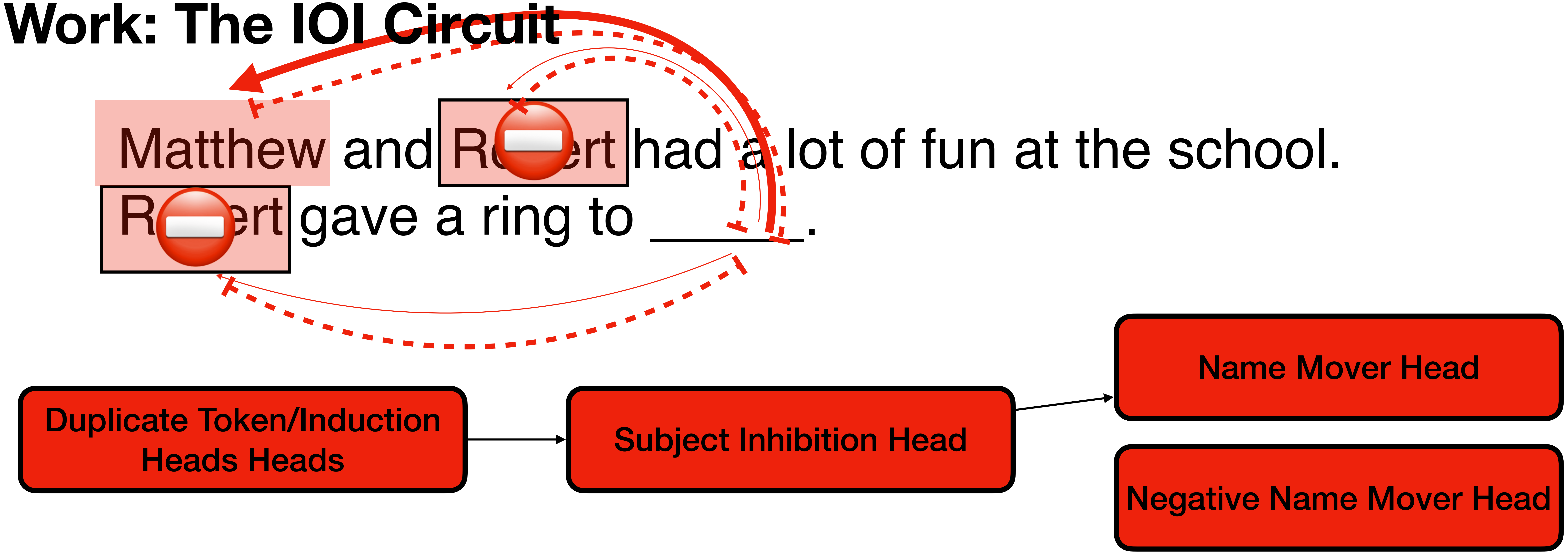
Prior Work: The IOI Circuit

Matthew and Robert had a lot of fun at the school.
Robert gave a ring to _____.

1. Identify any duplicated names.
2. Alert the S-Inhibition head of their location
3. Block attention to these duplicates

Understanding LLM Circuits and Algorithms

Prior Work: The IOI Circuit



1. Identify any duplicated names.
2. Alert the S-Inhibition head of their location
3. Block attention to these duplicates
4. Attend to remaining names and copy

Understanding LLM Circuits and Algorithms

Prior Work: The IOI Circuit

Matthew and Robert had a lot of fun at the school.
Robert gave a ring to **Matthew**.

1. Identify any duplicated names.
2. Alert the S-Inhibition head of their location
3. Block attention to these duplicates
4. Attend to remaining names and copy

Understanding LLM Circuits and Algorithms

Generalization to the Colored Objects Circuit

Q: On the table, I see an orange textbook, a red puzzle, and a purple cup. What color is the textbook?
A: Orange
Q: One the table, there is a blue pencil, a black necklace, and a yellow lighter. What color is the pencil?
A: _____

Understanding LLM Circuits and Algorithms

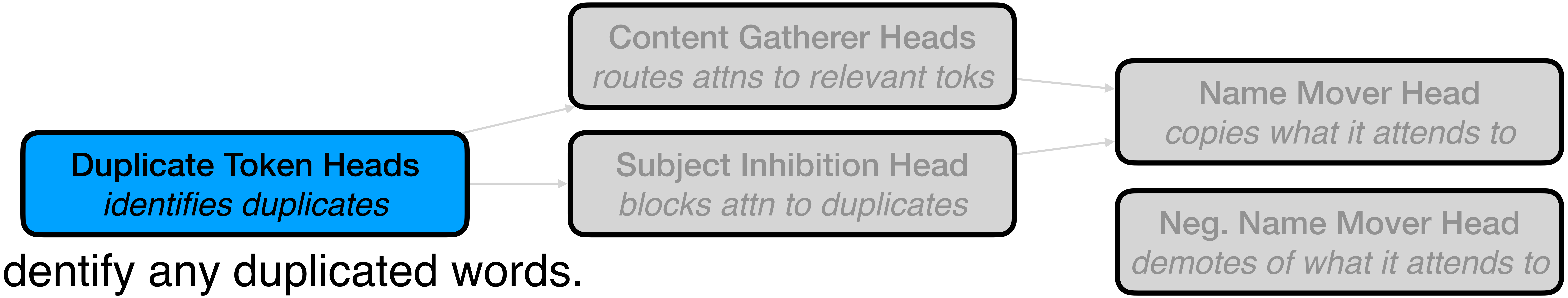
Generalization to the Colored Objects Circuit

Q: On the table, I see an orange textbook, a red puzzle, and a purple cup. What color is the textbook?
A: Orange
Q: One the table, there is a blue pencil, a black necklace, and a yellow lighter. What color is the pencil?
A: _____

Understanding LLM Circuits and Algorithms

Generalization to the Colored Objects Circuit

Q: On the table, I see an orange textbook, a red puzzle, and a purple cup. What color is the textbook?
A: Orange
Q: One the table, there is a blue pencil, a black necklace, and a yellow lighter. What color is the pencil?
A: ______

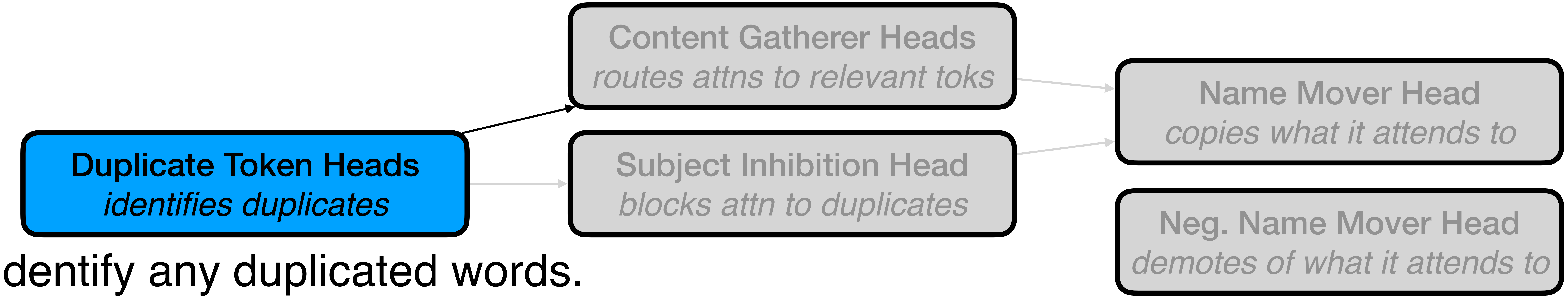


1. Identify any duplicated words.

Understanding LLM Circuits and Algorithms

Generalization to the Colored Objects Circuit

Q: On the table, I see an orange textbook, a red puzzle, and a purple cup. What color is the textbook?
A: Orange
Q: One the table, there is a blue pencil, a black necklace, and a yellow lighter. What color is the pencil?
A: ______

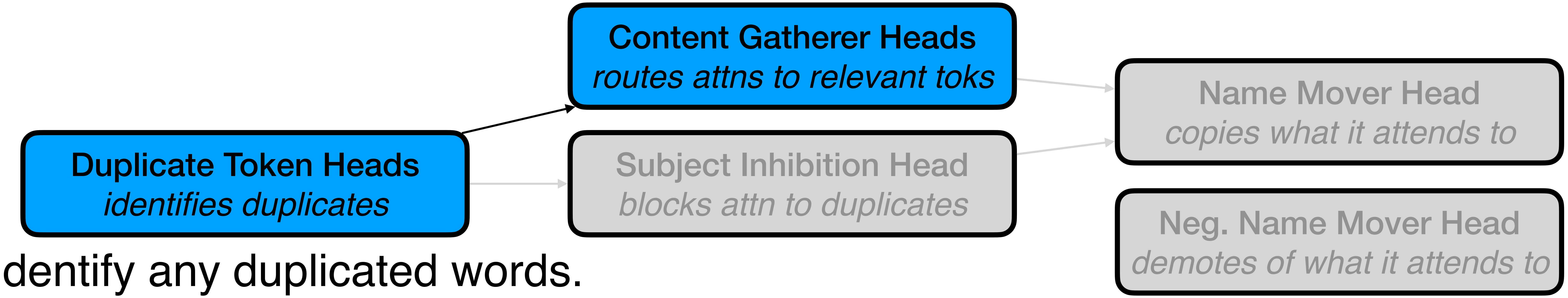


1. Identify any duplicated words.
2. Alert the Content Gatherer heads of their location

Understanding LLM Circuits and Algorithms

Generalization to the Colored Objects Circuit

Q: On the table, I see an orange textbook, a red puzzle, and a purple cup. What color is the textbook?
A: Orange
Q: One the table, there is a blue pencil, a black necklace, and a yellow lighter. What color is the pencil?
A: ______



1. Identify any duplicated words.
2. Alert the Content Gatherer heads of their location
3. Promote attention to these duplicates

Understanding LLM Circuits and Algorithms

Generalization to the Colored Objects Circuit

Q: On the table, I see an orange textbook, a red puzzle, and a purple cup. What color is the textbook?
A: Orange
Q: One the table, there is a blue pencil, a black necklace, and a yellow lighter. What color is the pencil?
A: ______

Content Gatherer Heads
routes attns to relevant toks

Name Mover Head
copies what it attends to

Duplicate Token Heads
identifies duplicates

Subject Inhibition Head
blocks attn to duplicates

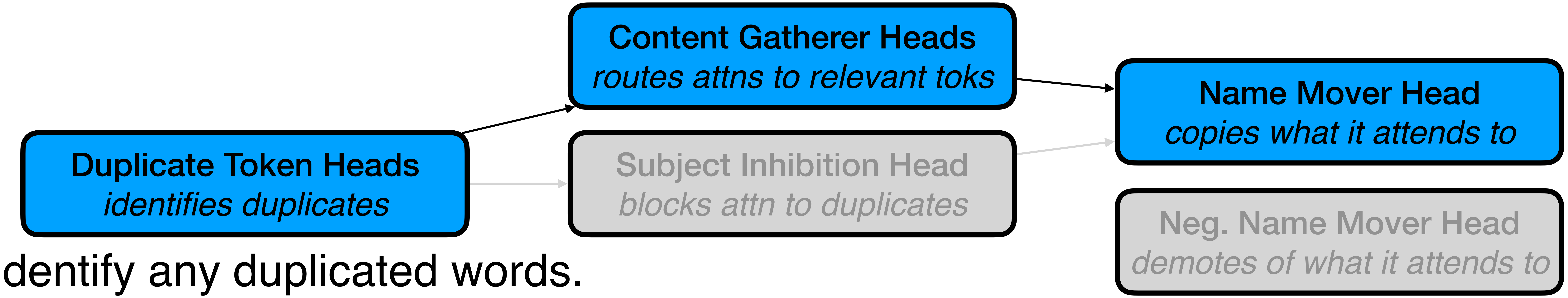
Neg. Name Mover Head
demotes of what it attends to

1. Identify any duplicated words.
2. Alert the Content Gatherer heads of their location
3. Promote attention to these duplicates
4. Attend to (color of) duplicate and copy

Understanding LLM Circuits and Algorithms

Generalization to the Colored Objects Circuit

Q: On the table, I see an orange textbook, a red puzzle, and a purple cup. What color is the textbook?
A: Orange
Q: One the table, there is a blue pencil, a black necklace, and a yellow lighter. What color is the pencil?
A: **Blue**



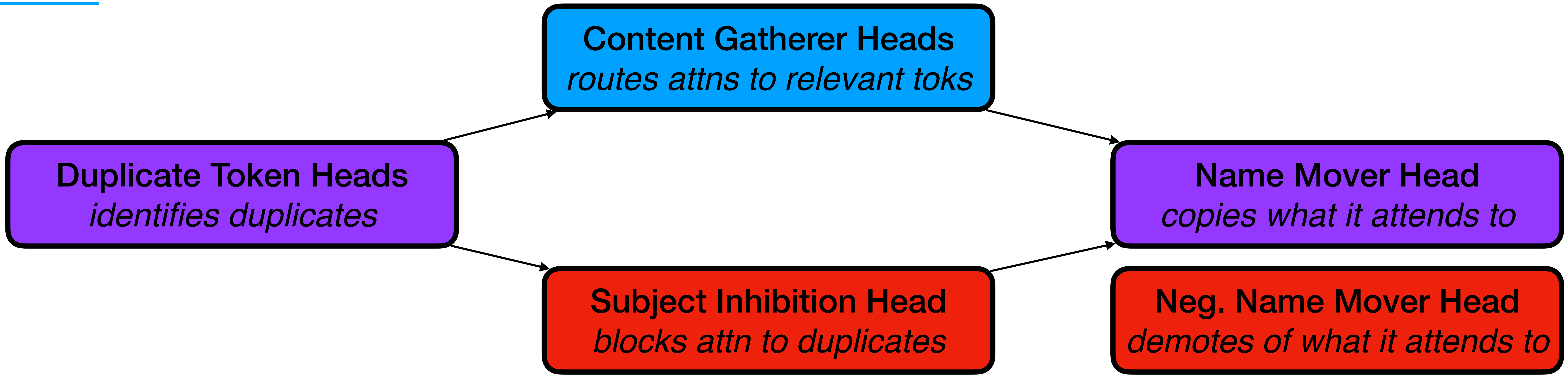
1. Identify any duplicated words.
2. Alert the Content Gatherer heads of their location
3. Promote attention to these duplicates
4. Attend to (color of) duplicate and copy

Understanding LLM Circuits and Algorithms

Circuit Similarities and Differences

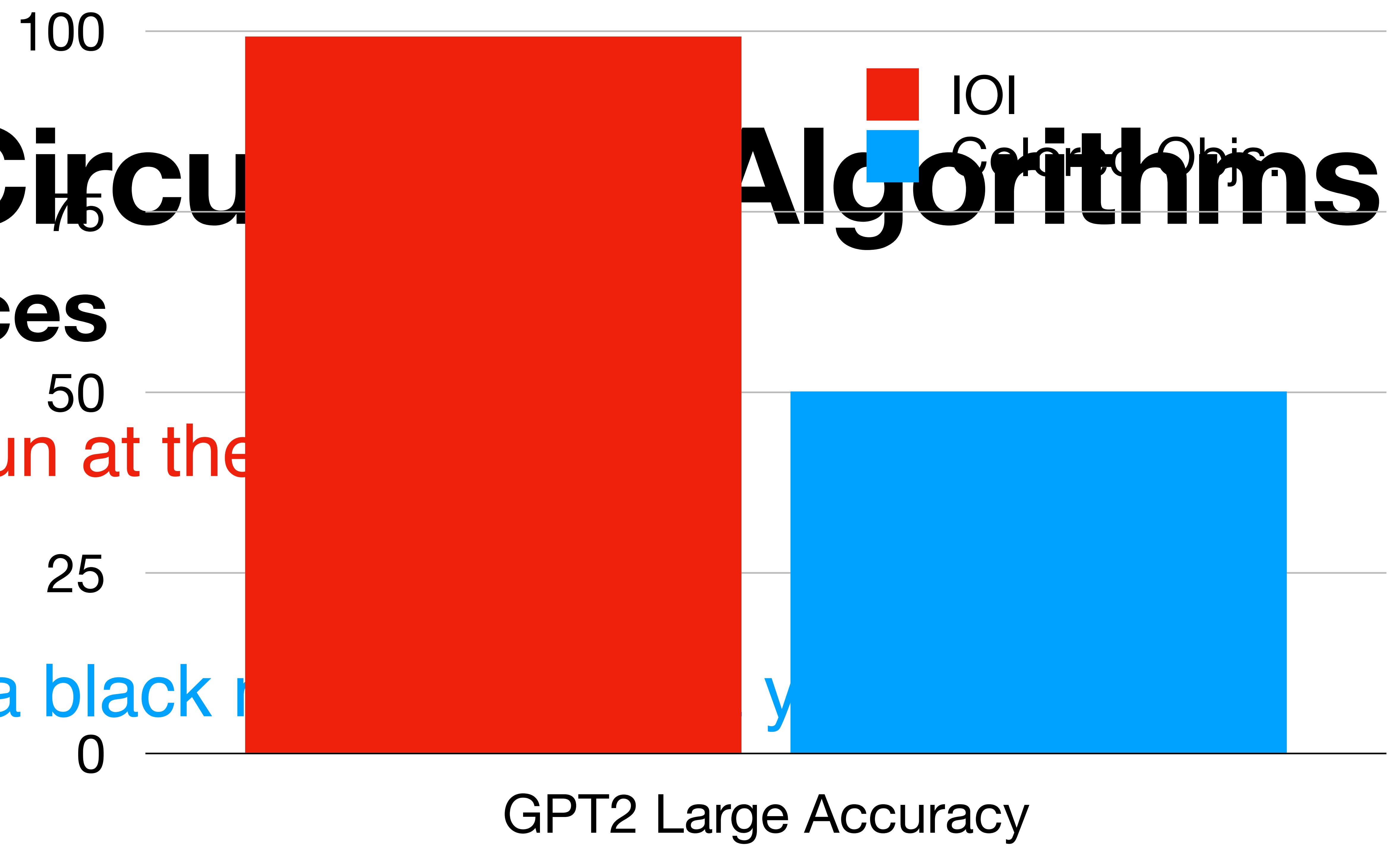
Then, Matthew and Robert had a lot of fun at the school.
Robert gave a ring to _____

Q: One the table, there is a blue pencil, a black necklace, and a yellow lighter.
What color is the pencil?
A: _____



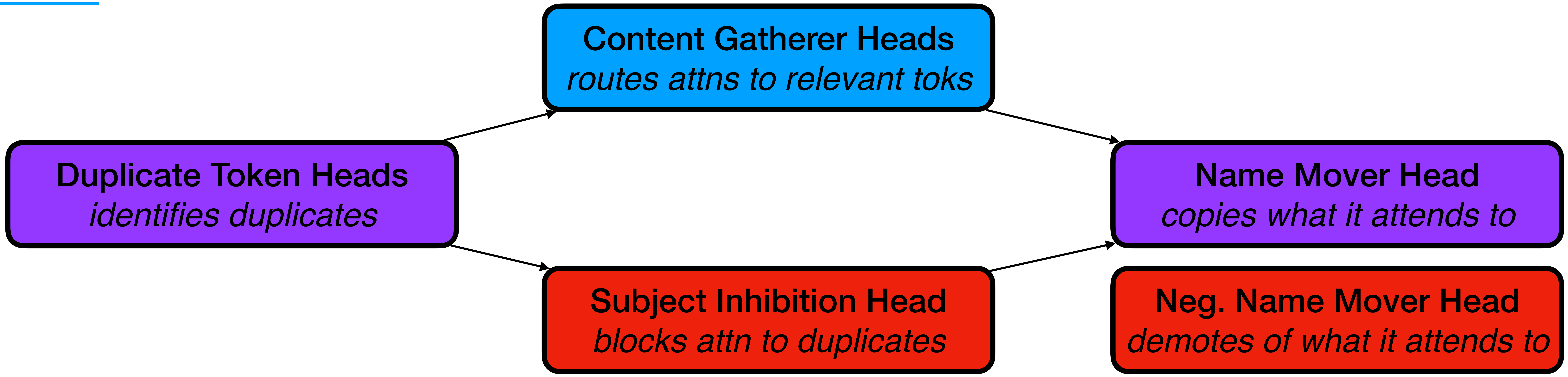
Understanding LLM Circu Algorithms

Circuit Similarities and Differences



Then, Matthew and Robert had a lot of fun at the
Robert gave a ring to _____

Q: One the table, there is a blue pencil, a black
What color is the pencil?
A: _____

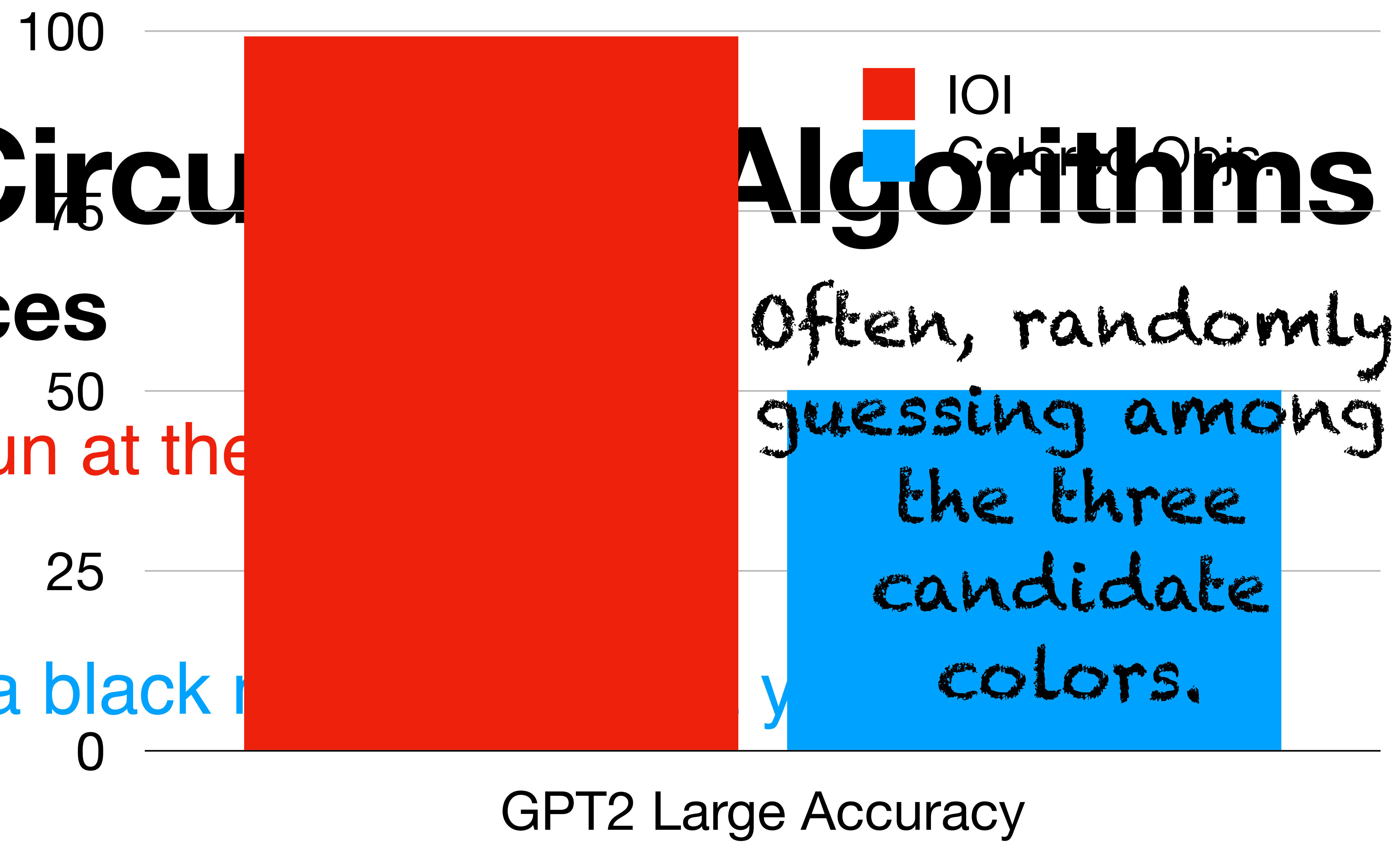


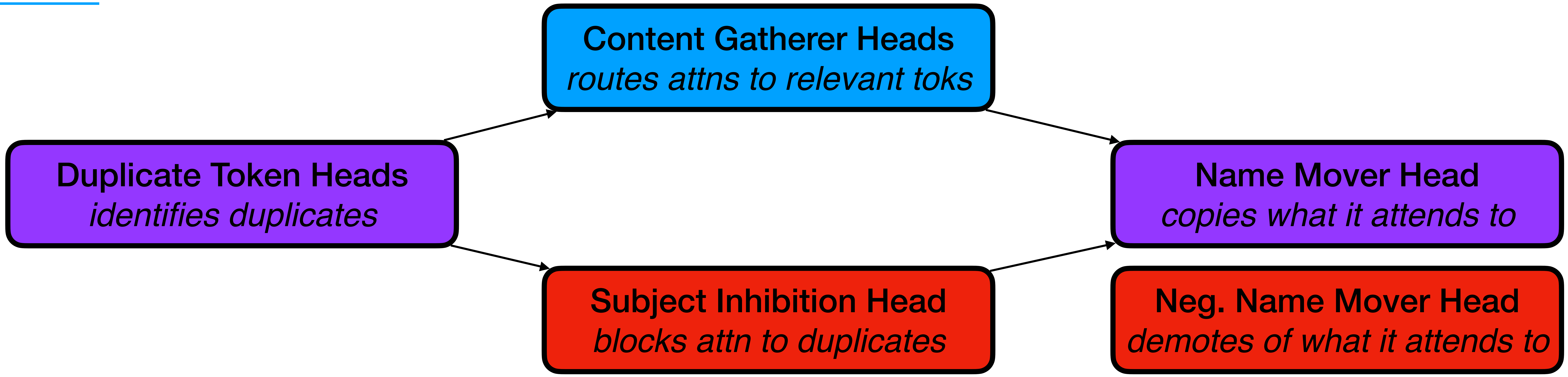
Understanding LLM Circu Algorithms

Circuit Similarities and Differences

Then, Matthew and Robert had a lot of fun at the
Robert gave a ring to _____

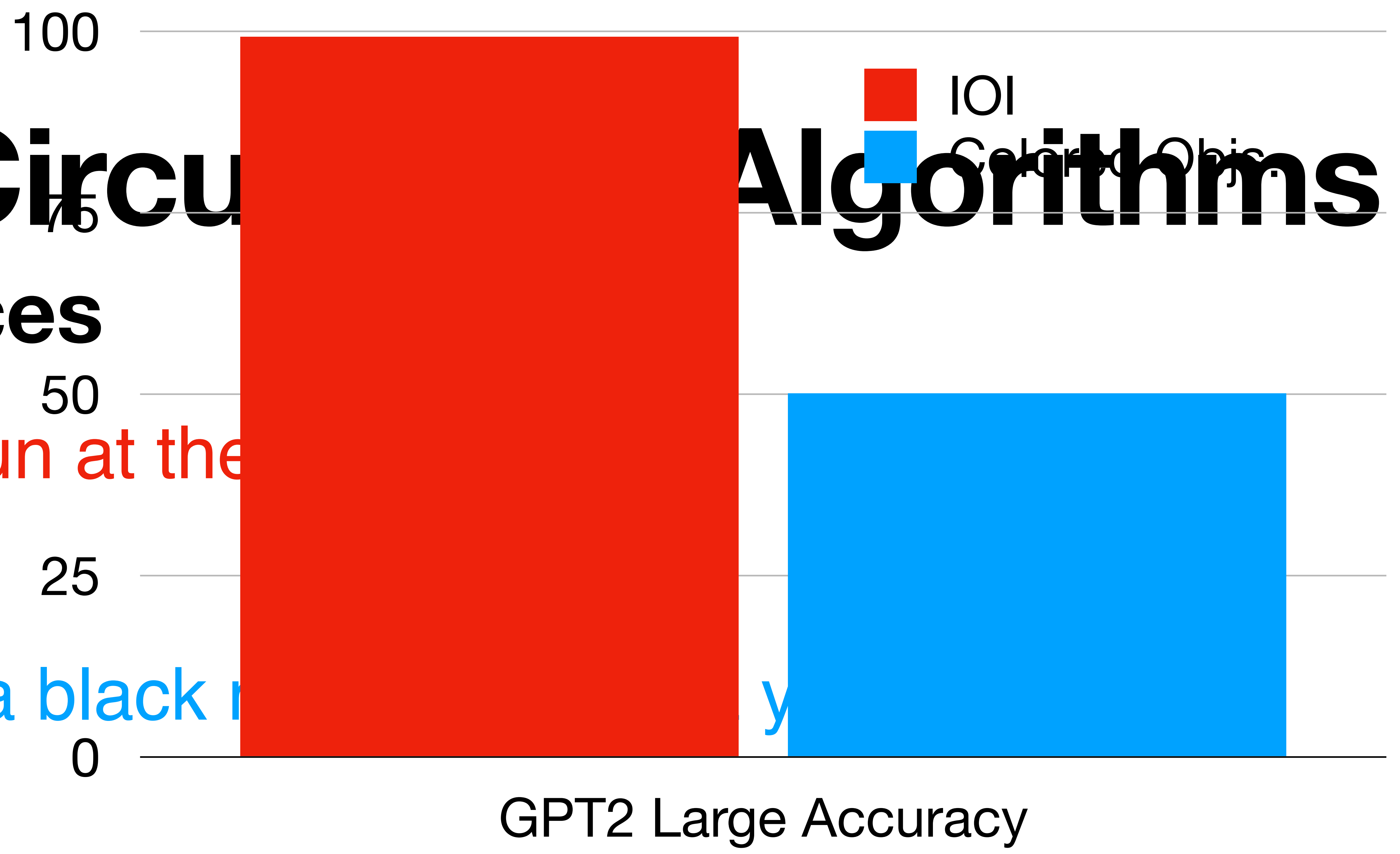
Q: One the table, there is a blue pencil, a black r y
What color is the pencil?
A: _____





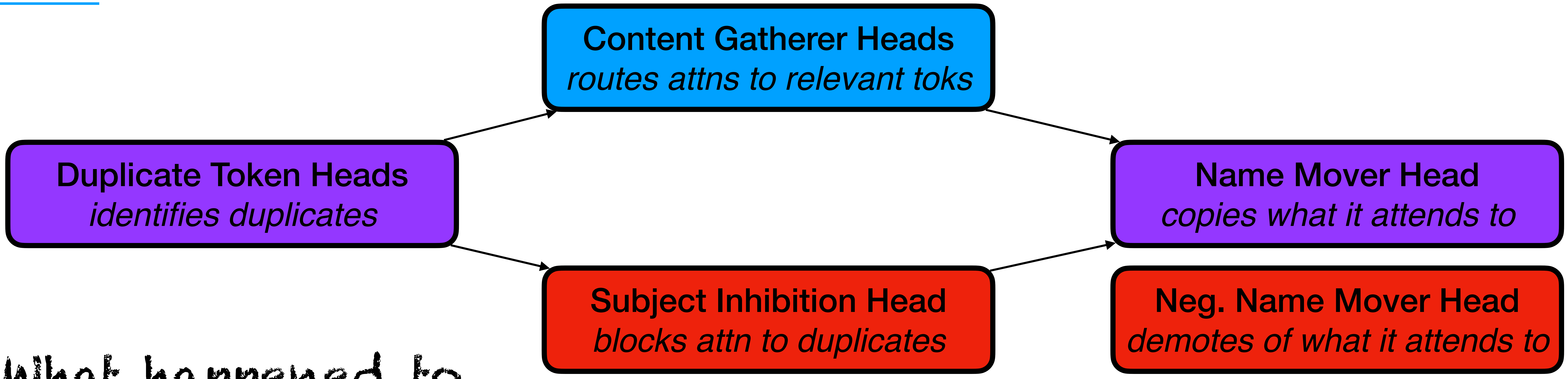
Understanding LLM Circu… Algorithms

Circuit Similarities and Differences



Then, Matthew and Robert had a lot of fun at the…
Robert gave a ring to _____

Q: One the table, there is a blue pencil, a black …
What color is the pencil?
A: _____

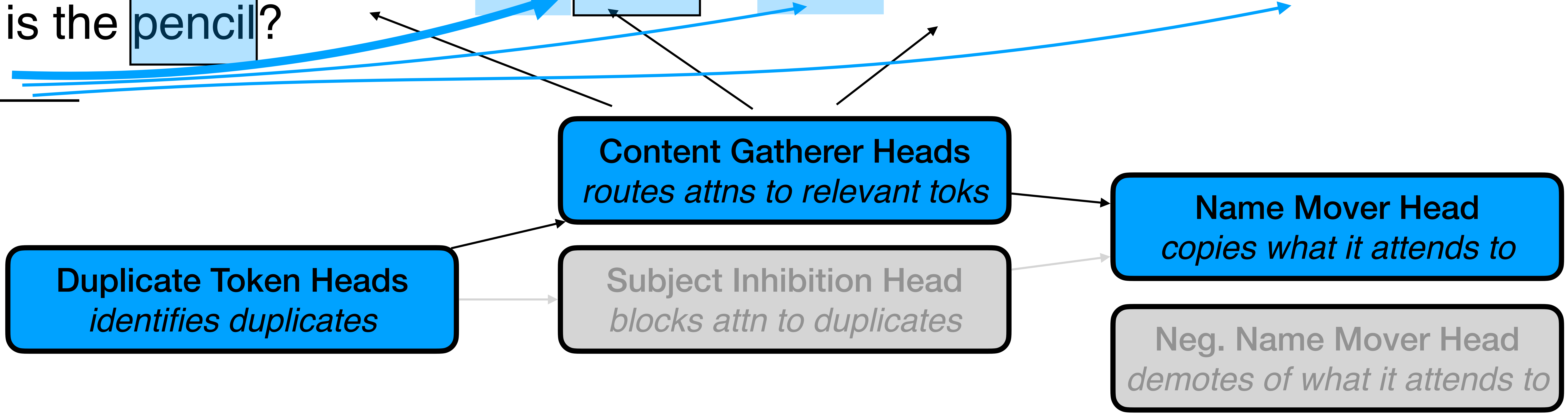


What happened to
the inhibition?

Understanding LLM Circuits and Algorithms

Generalization to the Colored Objects Circuit

Q: On the table, I see an orange textbook, a red puzzle, and a purple cup. What color is the textbook?
A: Orange
Q: One the table, there is a blue pencil, a black necklace, and a yellow lighter. What color is the pencil?
A: ______



Understanding LLM Circuits and Algorithms

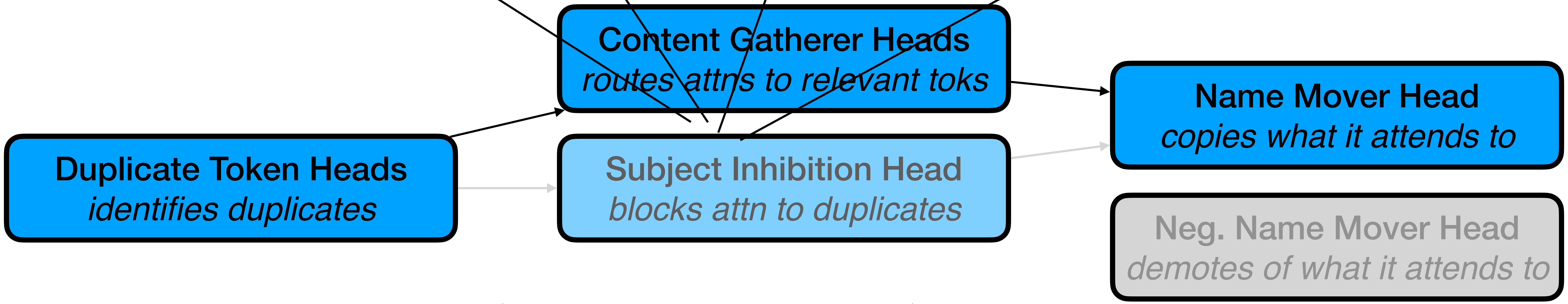
Generalization to the Colored Objects Circuit

Q: On the table, I see an orange textbook, a red puzzle, and a purple cup. What color is the textbook?

A: Orange

Q: One the table, there is a blue pencil, a black necklace, and a yellow lighter. What color is the pencil?

A: ______



Active, but incorrect biases
and weak signal

Understanding LLM Circuits and Algorithms

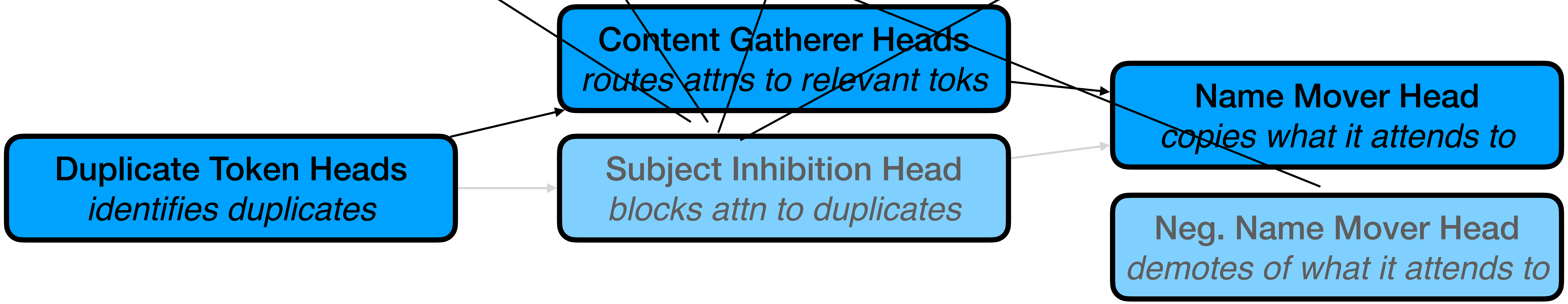
Generalization to the Colored Objects Circuit

Q: On the table, I see an orange textbook, a red puzzle, and a purple cup. What color is the textbook?

A: Orange

Q: One the table, there is a blue pencil, a black necklace, and a yellow lighter. What color is the pencil?

A: ______



Functioning as expected,
but "benched"

Understanding LLM Circuits and Algorithms

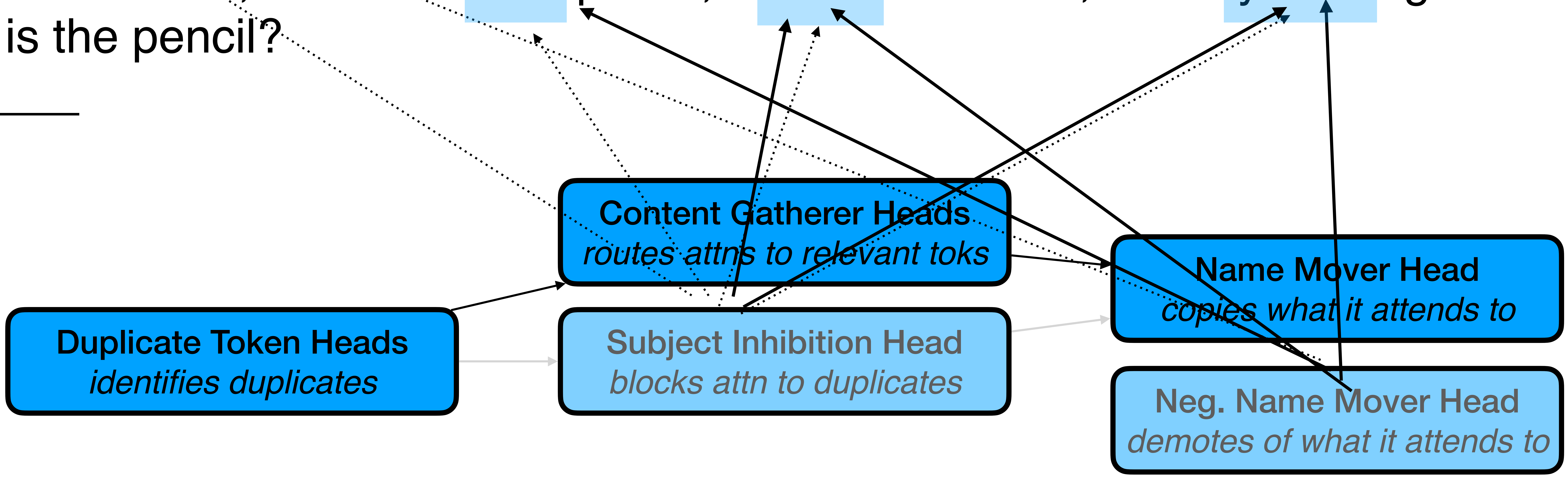
Generalization to the Colored Objects Circuit

Q: On the table, I see an orange textbook, a red puzzle, and a purple cup. What color is the textbook?

A: Orange

Q: One the table, there is a blue pencil, a black necklace, and a yellow lighter. What color is the pencil?

A: ______



Intervene to force the desired attention pattern (based on IOI) for just these 4 heads.

Ur… ms

Ge…

Q: C… …t

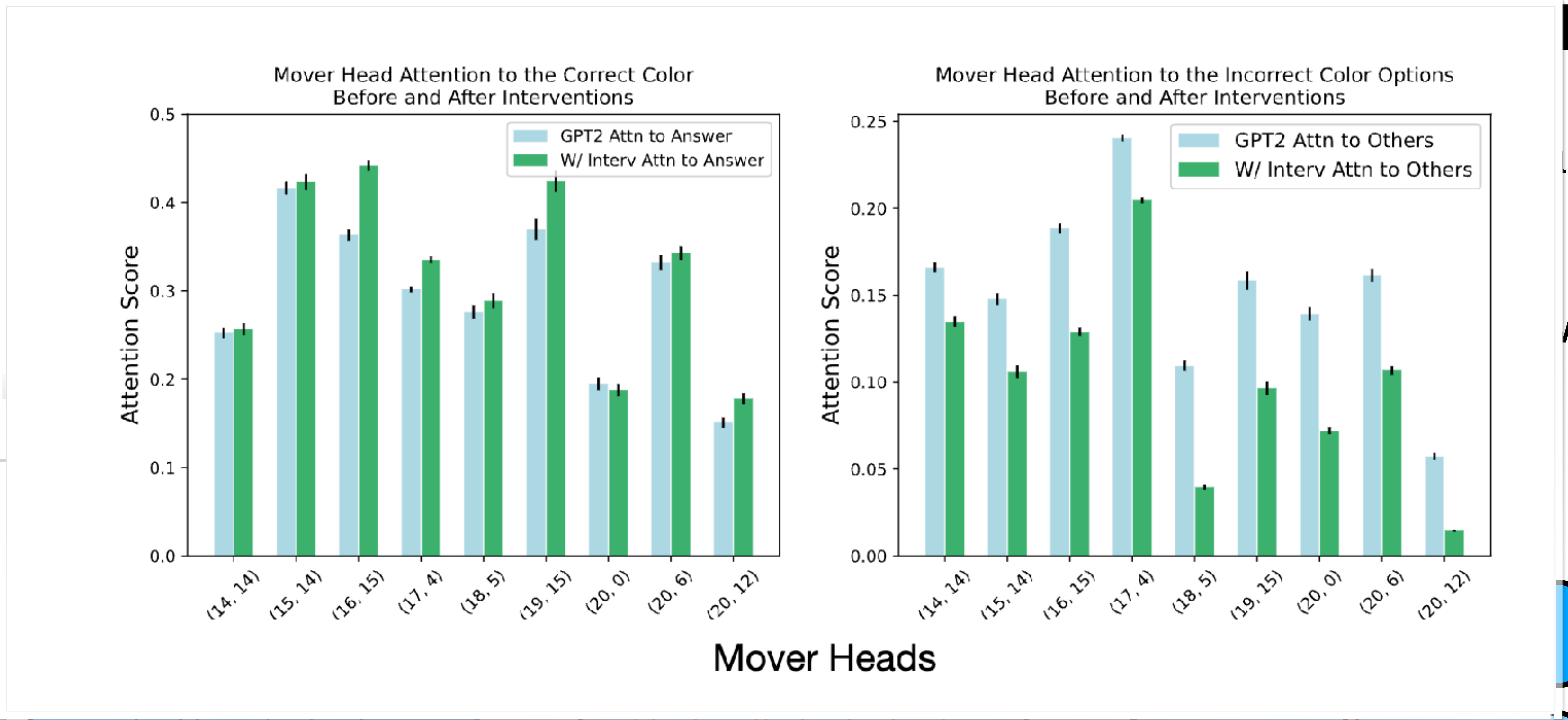
colo…

A: C…

Q: C… …What

colo…

A: _



identifies duplicates

blocks attn to duplicates

Neg. Name Mover Head
demotes of what it attends to

Intervene to force the desired attention pattern (based on IOI) for just these 4 heads.

Understanding LLM Circuits and Algorithms

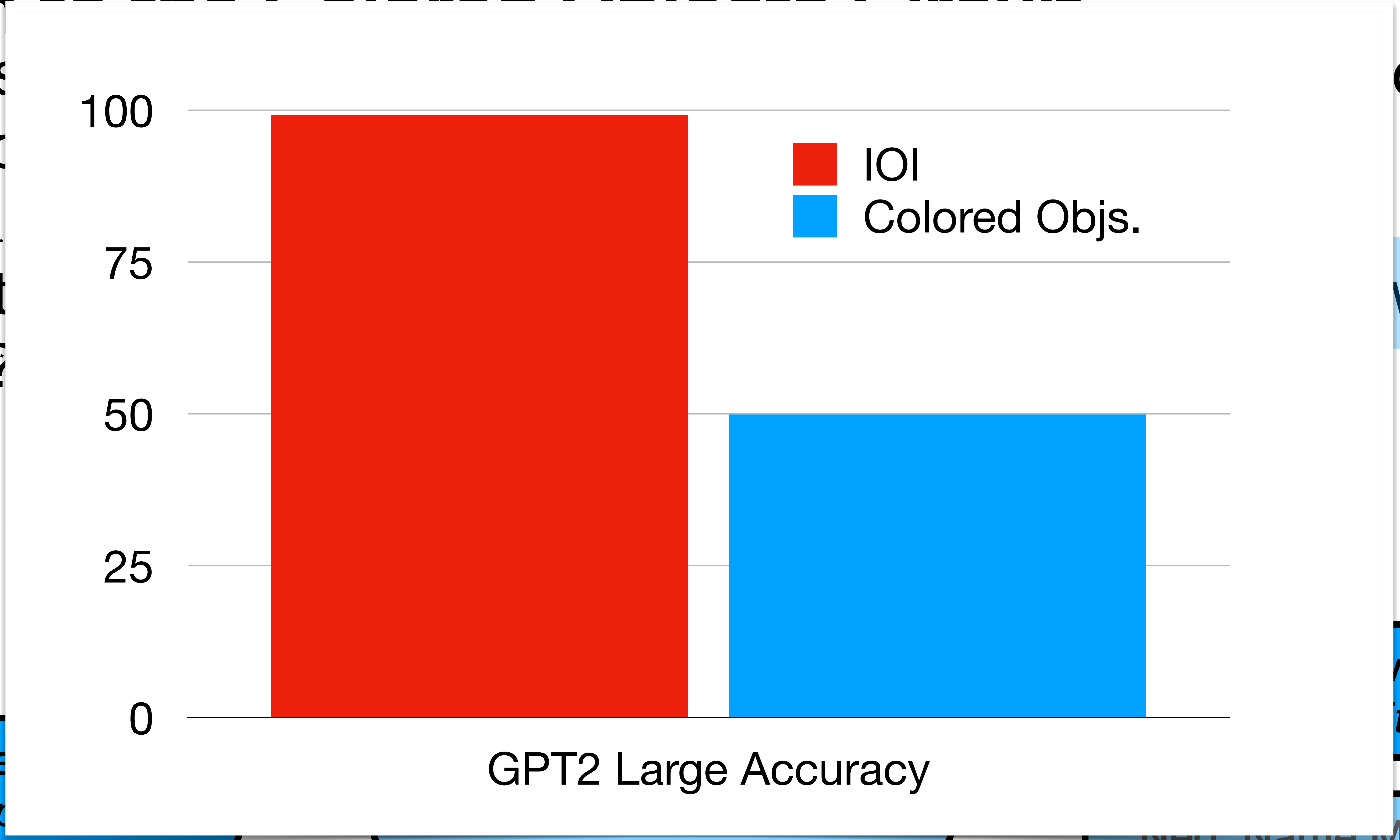
Generalization to the Colored Objects Circuit

Q: On the table, I s... cup. What color is the textboo...

A: Orange

Q: One the table, t... lighter. What color is the pencil?

A: ______



Duplicate Toke...
identifies dup...

...ver Head
...t attends to

Neg. Name Mover Head
demotes of what it attends to

Intervene to force the desired attention pattern (based on IOI) for just these 4 heads.

Understanding LLM Circuits and Algorithms

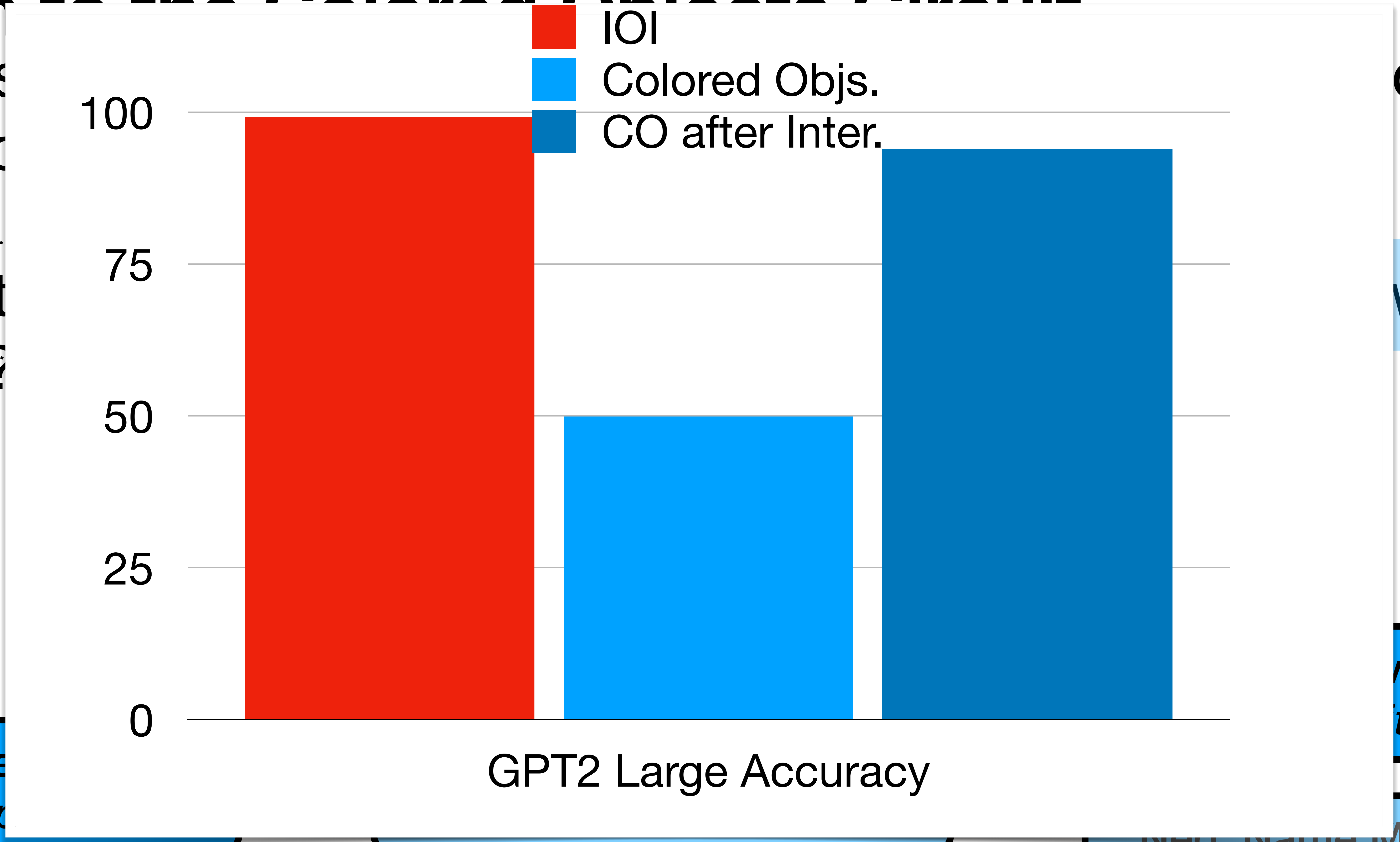
Generalization

Q: On the table, I s cup. What color is the textboo

A: Orange

Q: One the table, t v lighter. What color is the pencil?

A: ______



Duplicate Toke
identifies dup

ver Head
t attends to

Neg. Name Mover Head
demotes of what it attends to

Intervene to force the desired attention pattern (based on IOI) for just these 4 heads.

Understanding LLM Circuits and Algorithms

Summary and Discussion

- There is evidence that individual circuit components can be modular and generic, and reused across tasks
- This reuse gives us insight into the algorithmic “building blocks” of Transformers, which might not match our intuitions (e.g., from linguistics) about how tasks decompose into subtasks, and which can explain otherwise arbitrary-seeming behaviors like sensitivity to prompts
- Mending a “broken” circuit can have substantial effects on performance
- Follow up work in progress:
    - Why doesn’t the LLM learn to the correct circuit itself (hypothesis: undertrained/effect of scale/grokking)
    - Similarities to human neural mechanisms — (emergent) capacity limits, chunking, primacy/recency biases, content effects, curriculum effects….

Discussion

- LLMs are often assumed to be black boxes. They aren't.
- Interpretting LLMs in higher-level functional terms can offer insight into the "neurocircuitry" and "cognition" of LLMs…
- …which might substantially transform future work in theory, engineering, safety, and even the science of human language and cognition
- But its a long game! So much still unknown:
  - Methods are new and primitive. We cannot take results for granted.
  - We don't know what we are looking for, or have good metrics of success.
  - Moving targets. Models keep changing, and interpretability results don't always generalize
- But problems that are long-term and challenging are good things for scientists! Lots of reasons to be excited and optimistic :)

Thank you!

Jack Merullo

Qinan Yu

Carsten Eickhoff