Large Language Models as Cognitive Models

Neuroscience Forum Workshop: Exploring the Bidirectional Relationship Between Artificial Intelligence and Neuroscience

BROWN

Ellie Pavlick, March 25 2024

Disclosures

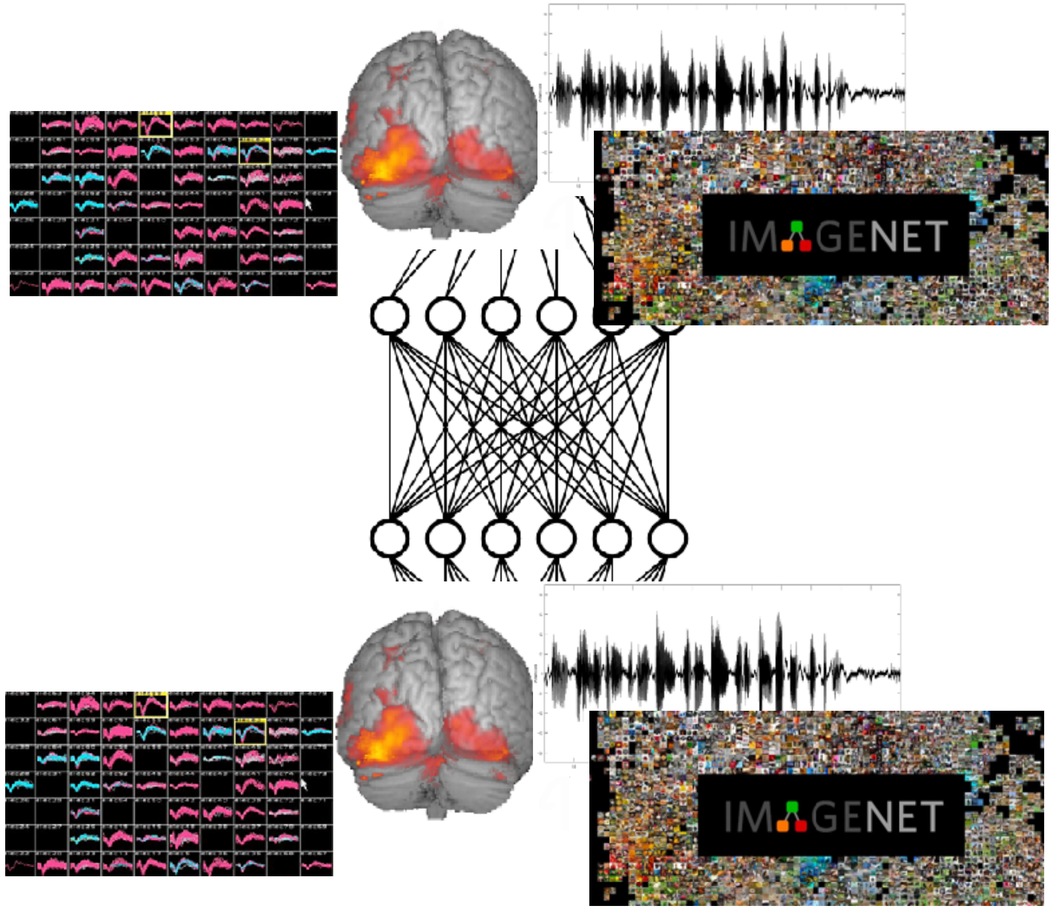
- Current/Recent Research Support: NSF, DARPA, IARPA, Google, ONR, NIH
- Consulting/Speaking Engagements: Google Deepmind (Current Employee); Signify (One-Time Consultant)

The Role of AI in Cognitive Neuroscience

Predictive Tool	Explanatory Model

The Role of AI in Cognitive Neuroscience

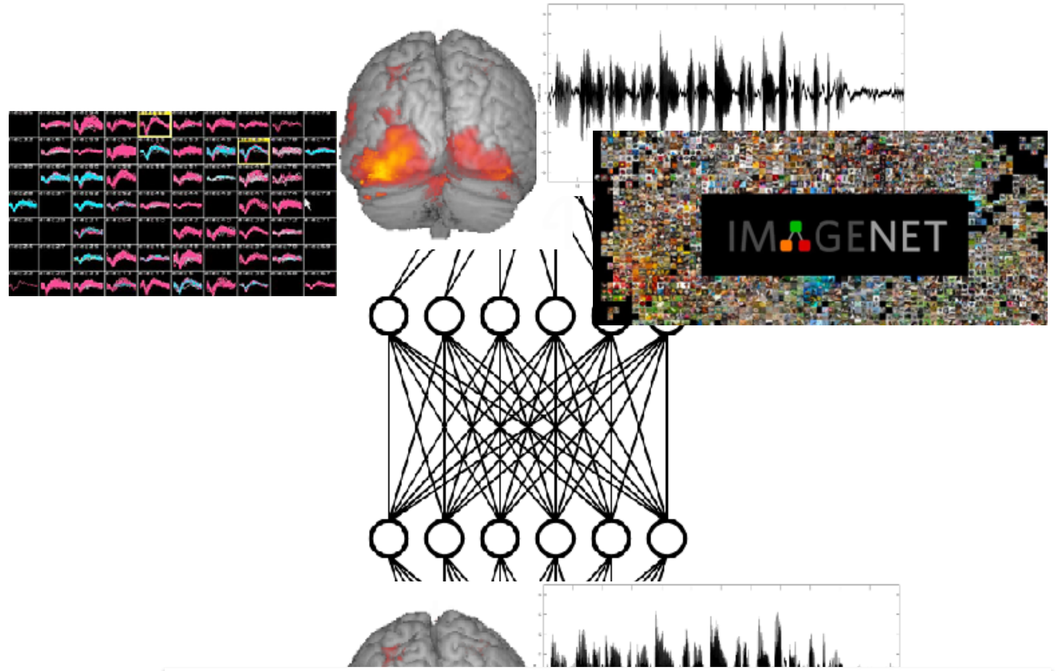
Predictive Tool



Explanatory Model

The Role of AI in Cognitive Neuroscience

Predictive Tool



e.g.,

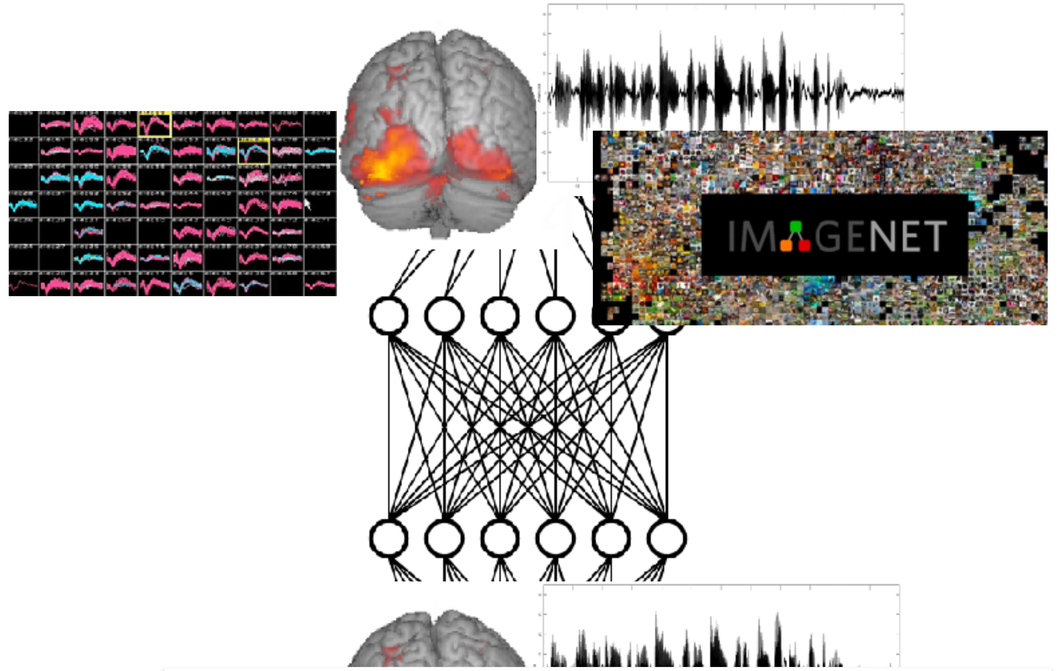
BRAINGATE

TURNING THOUGHT INTO ACTION

Explanatory Model

The Role of AI in Cognitive Neuroscience

Predictive Tool

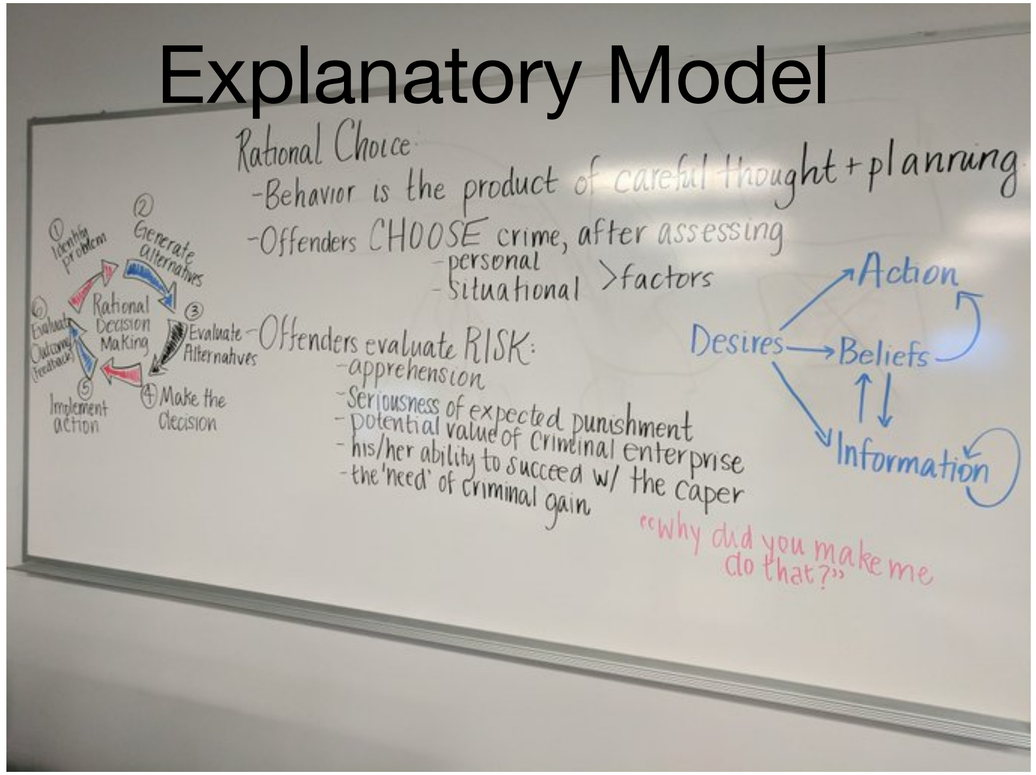


e.g.,

BRAINGATE

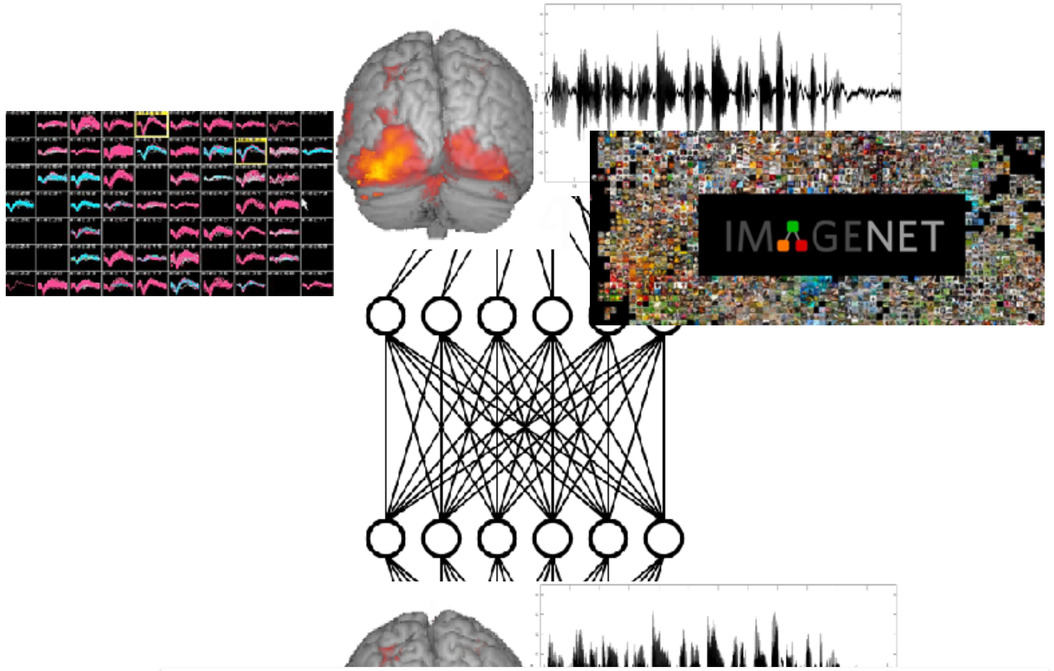
TURNING THOUGHT INTO ACTION

Explanatory Model



The Role of AI in Cognitive Neuroscience

Predictive Tool

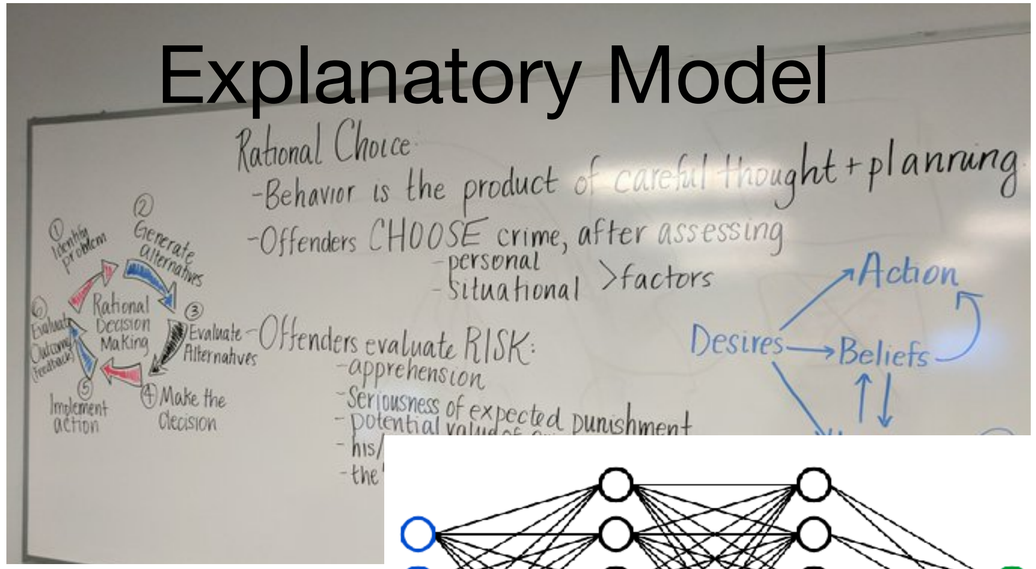


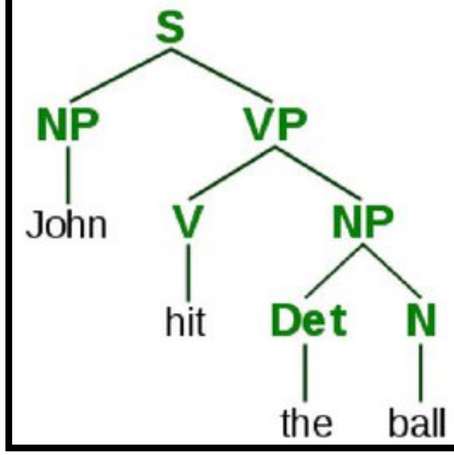
e.g.,

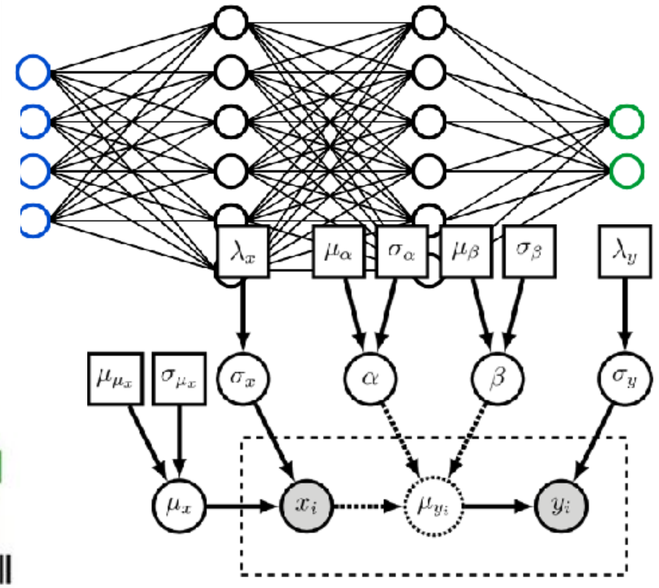
BRAINGATE

TURNING THOUGHT INTO ACTION

Explanatory Model







The Role of AI in Cognitive Neuroscience

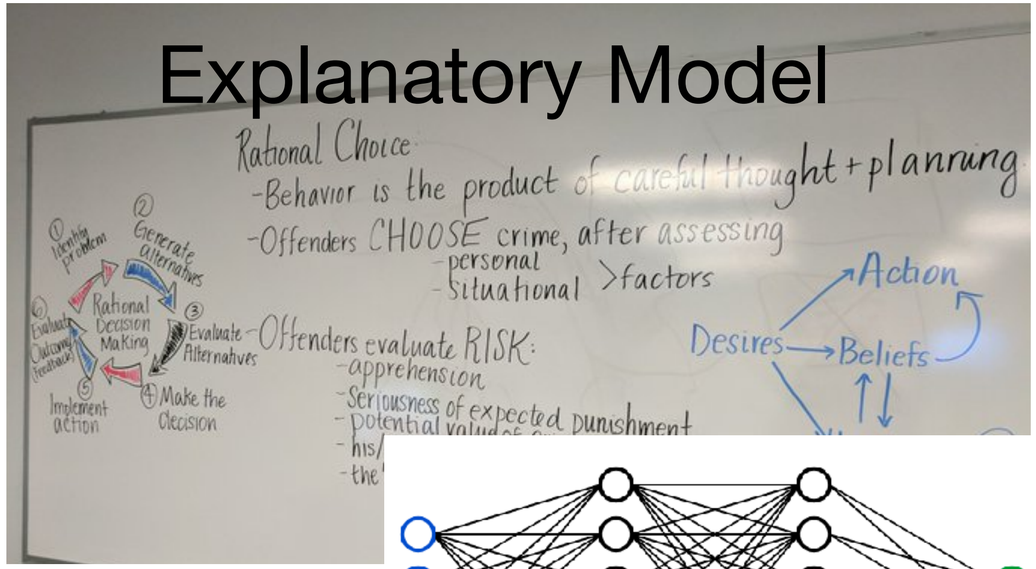
Predictive Tool

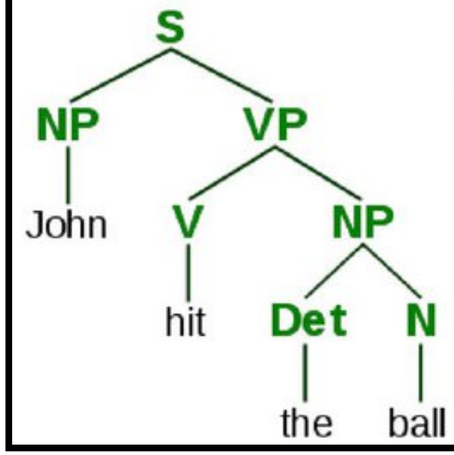
e.g.,

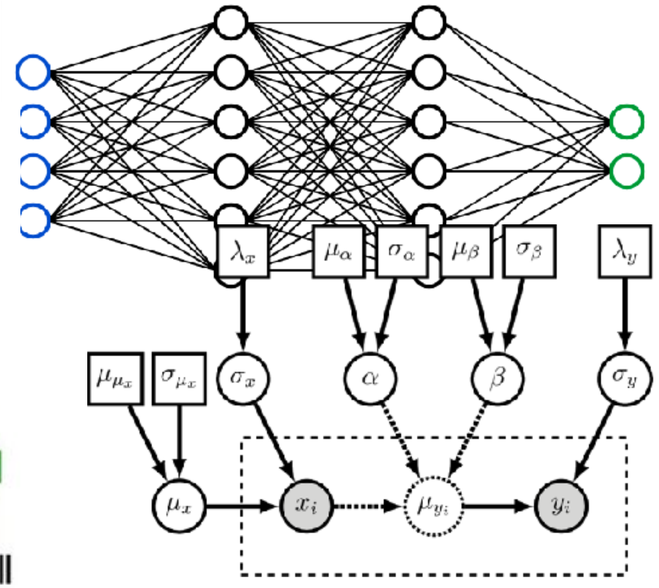
BRAINGATE

TURNING THOUGHT INTO ACTION

Explanatory Model







Reasons *not* to use LLMs as Cognitive Models

Reasons ***not*** to use LLMs as Cognitive Models

LLMs were not designed as cognitive models.

The success was "stumbled upon" via good engineering, but they do not represent any coherent theory. And, they still are woefully un-human-like in so many ways.

Reasons *not* to use LLMs as Cognitive Models

LLMs were not designed as cognitive models.

The success was "stumbled upon" via good engineering, but they do not represent any coherent theory. And, they still are woefully un-human-like in so many ways.

LLMs are black boxes.

We do not understand how they work. We cannot simply replace one black box—i.e., the human brain and mind—with another black box—i.e, the artificial neural network.

Reasons *to* use LLMs as Cognitive Models

LLMs were not designed as cognitive models.

Yes. But their ability to simulate human behavior across so many varied domains is worth taking seriously. And in certain some cases (namely, language) they represent our best computational model of human behavior by a long shot.

LLMs are black boxes.

Reasons *to* use LLMs as Cognitive Models

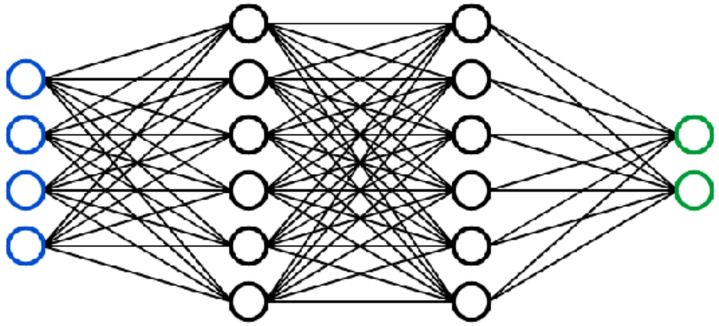
LLMs were not designed as cognitive models.

Yes. But their ability to simulate human behavior across so many varied domains is worth taking seriously. And in certain some cases (namely, language) they represent our best computational model of human behavior by a long shot.

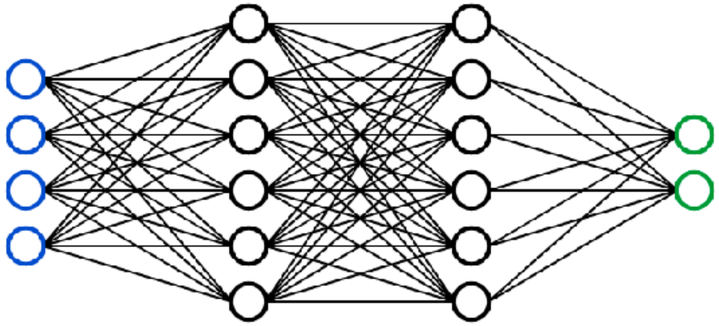
LLMs are black boxes.

Not necessarily. In fact, recent work shows LLMs might be more interpretable than we often assume, and uncovering these mechanisms could form a virtuous cycle that advances cognitive neuroscience+AI in tandem.

Virtuous Cycle between (Generative) AI and Cognitive Neuroscience

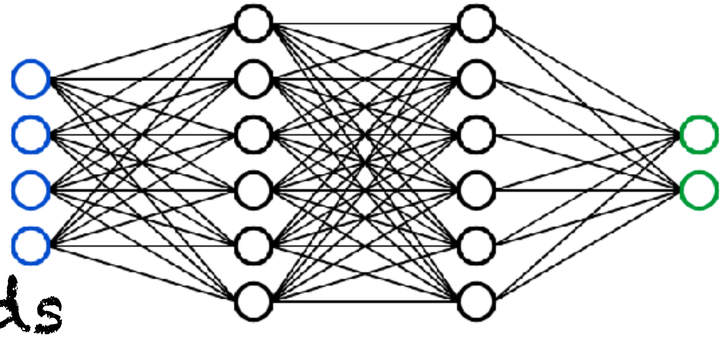


Virtuous Cycle between (Generative) AI and Cognitive Neuroscience



Hypotheses about representations,
structures, mechanisms...

Virtuous Cycle between (Generative) AI and Cognitive Neuroscience

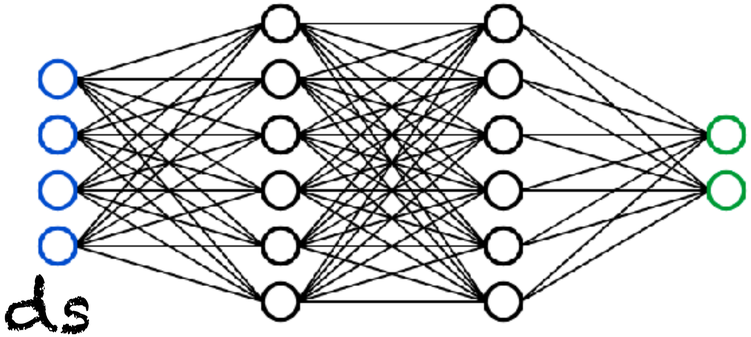


Analytical methods
(relational similarity,
manifolds...)

Hypotheses about representations,
structures, mechanisms...

Virtuous Cycle between (Generative) AI and Cognitive Neuroscience

Hypotheses about representations, structures, mechanisms...



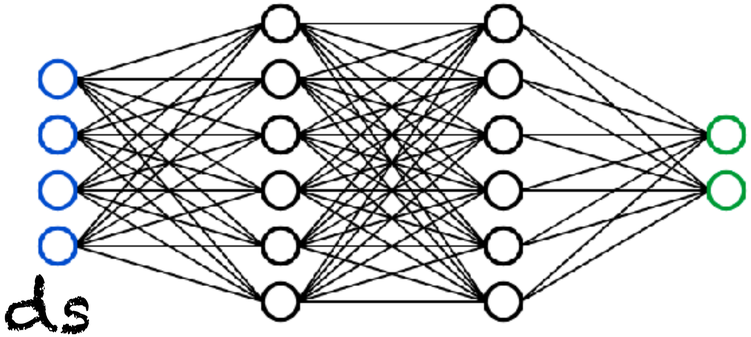
Analytical methods (relational similarity, manifolds...)

Hypotheses about representations, structures, mechanisms...

Virtuous Cycle between (Generative) AI and Cognitive Neuroscience

Hypotheses about representations, structures, mechanisms...

Forward-engineering, causal interventions ("lesions")



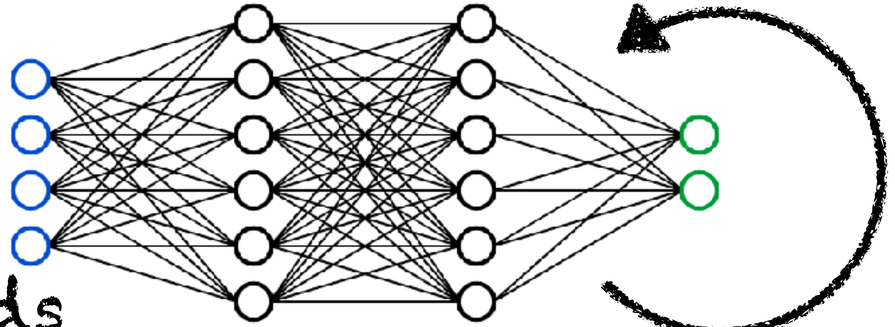
Analytical methods (relational similarity, manifolds...)

Hypotheses about representations, structures, mechanisms...

Virtuous Cycle between (Generative) AI and Cognitive Neuroscience

Hypotheses about representations, structures, mechanisms...

Forward-engineering, causal interventions ("lesions")



Analytical methods (relational similarity, manifolds...)

Improved generalization, sample efficiency, interpretability

Hypotheses about representations, structures, mechanisms...

Proofs of Concept

- LLMs are not merely “stochastic parrots”. They often contain **interpretable internal structure** which makes use of modular functionally-specialized components.
- These components can **sometimes resemble known brain mechanisms**, e.g., mechanisms for cognitive control in the prefrontal cortex
- LLMs, off the shelf, exhibit **behaviors for which we don’t have good competing computational cognitive models**, like the ability to learn flexibly both in context and in weights. Understanding these mechanisms and aligning them to those of humans could jointly offer new theories of human cognition and improve LLMs robustness, interpretability, and efficiency

Proofs of Concept

- LLMs are not merely “stochastic parrots”. They often contain **interpretable internal structure** which makes use of modular functionally-specialized components.
- These components can **sometimes resemble known brain mechanisms**, e.g., mechanisms for cognitive control in the prefrontal cortex
- LLMs, off the shelf, exhibit **behaviors for which we don’t have good competing computational cognitive models**, like the ability to learn flexibly both in context and in weights. Understanding these mechanisms and aligning them to those of humans could jointly offer new theories of human cognition and improve LLMs robustness, interpretability, and efficiency

Interpretable Internal Structure

What is the capital of France?

Paris

What is the capital of Poland?

Warsaw

Merullo, Jack, Carsten Eickhoff, and Ellie Pavlick. "Language models implement simple word2vec-style vector arithmetic." *NAACL* (2024).

Interpretable Internal Structure

Possibility #1: Idiosyncratic

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

Possibility #2: Systematic

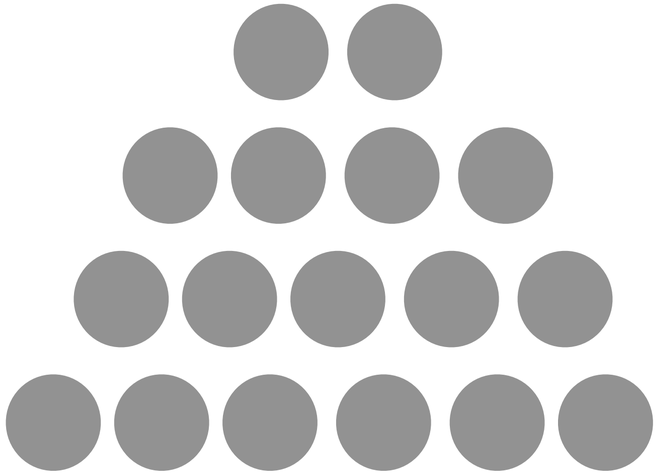
f|f(France) = Paris

f(Poland) = Warsaw

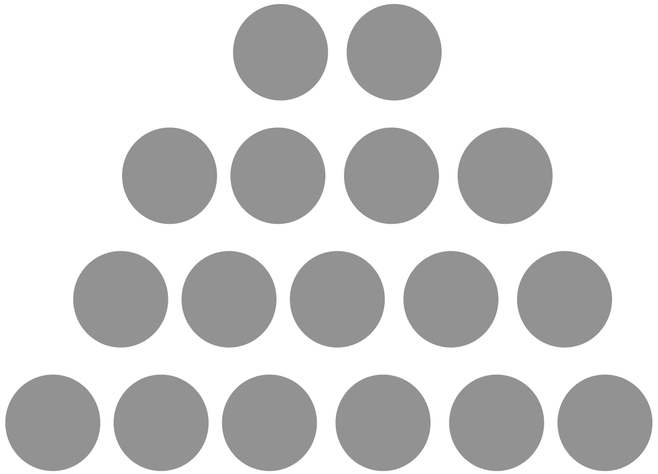
…ad… …ls implement simple word2vec-style vector arithmetic." *NAACL* (2024).

Interpretable Internal Structure

Warsaw



China

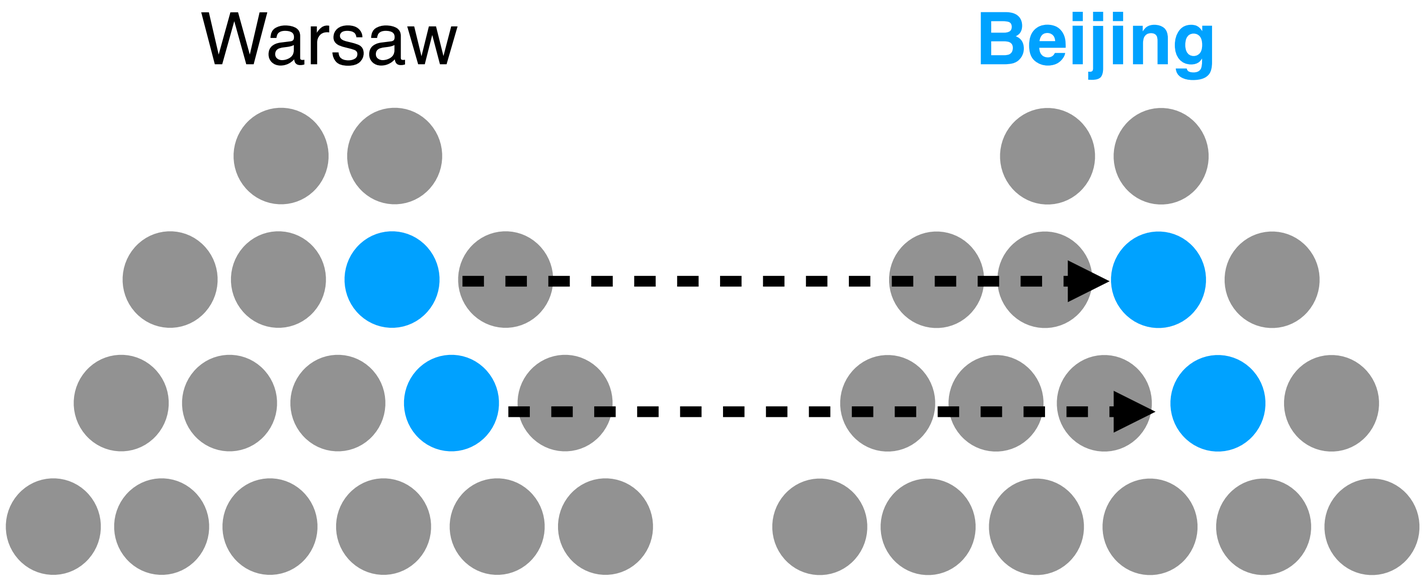


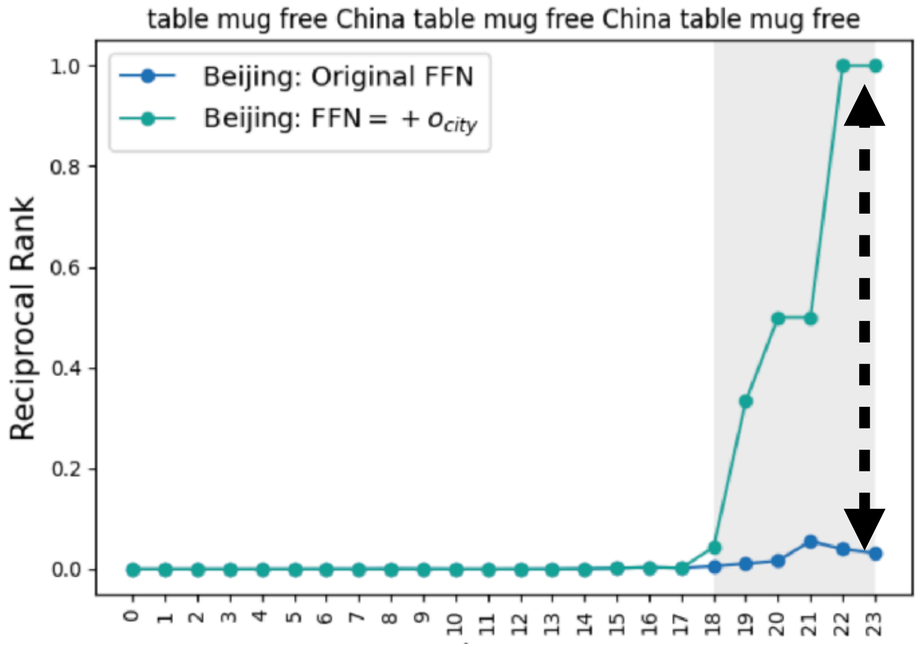
What is the capital of Poland?

beep boop China beep boop

Merullo, Jack, Carsten Eickhoff, and Ellie Pavlick. "Language models implement simple word2vec-style vector arithmetic." *NAACL* (2024).

Interpretable Internal Structure





What is the capital of Poland?

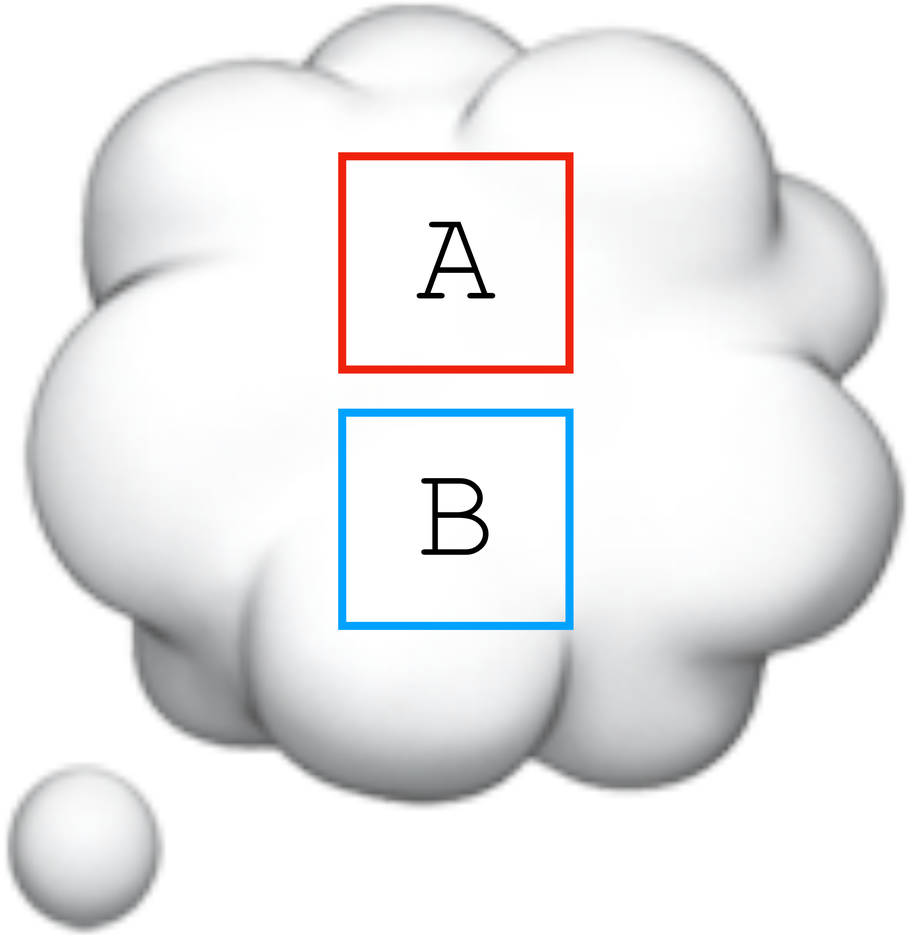
beep boop China beep boop

Merullo, Jack, Carsten Eickhoff, and Ellie Pavlick. "Language models implement simple word2vec-style vector arithmetic." *NAACL* (2024).

Proofs of Concept

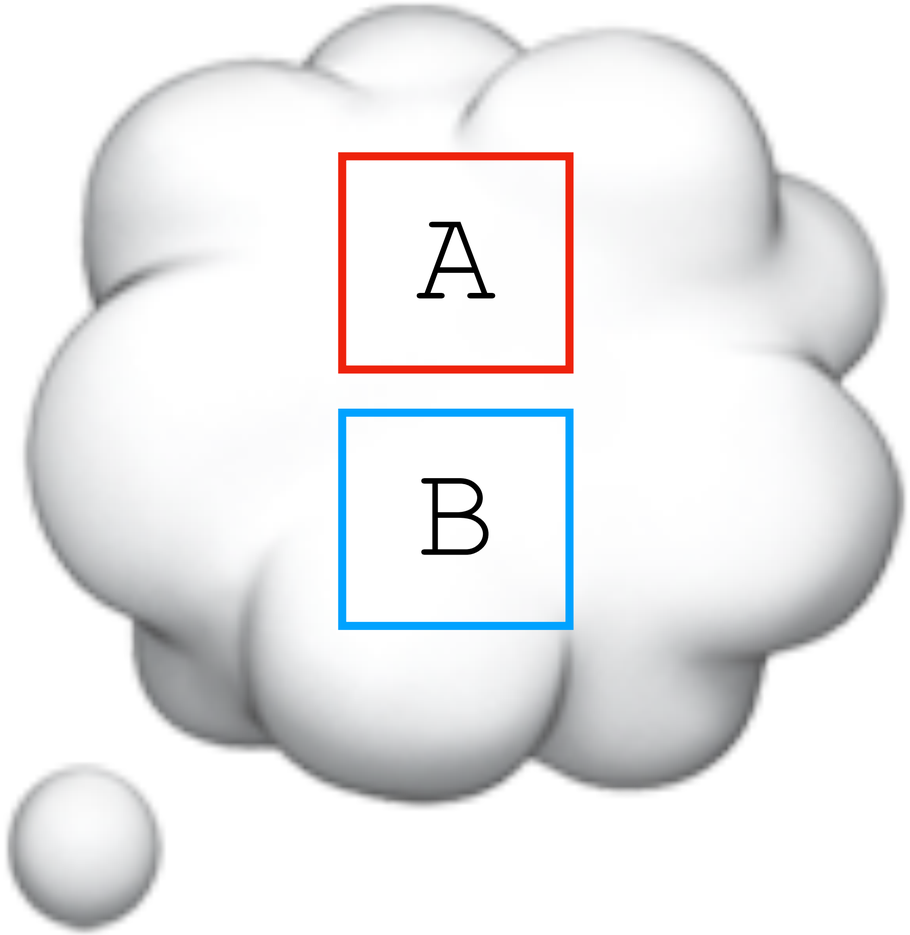
- LLMs are not merely “stochastic parrots”. They often contain **interpretable internal structure** which makes use of modular functionally-specialized components.
- These components can **sometimes resemble known brain mechanisms**, e.g., mechanisms for cognitive control in the prefrontal cortex
- LLMs, off the shelf, exhibit **behaviors for which we don’t have good competing computational cognitive models**, like the ability to learn flexibly both in context and in weights. Understanding these mechanisms and aligning them to those of humans could jointly offer new theories of human cognition and improve LLMs robustness, interpretability, and efficiency

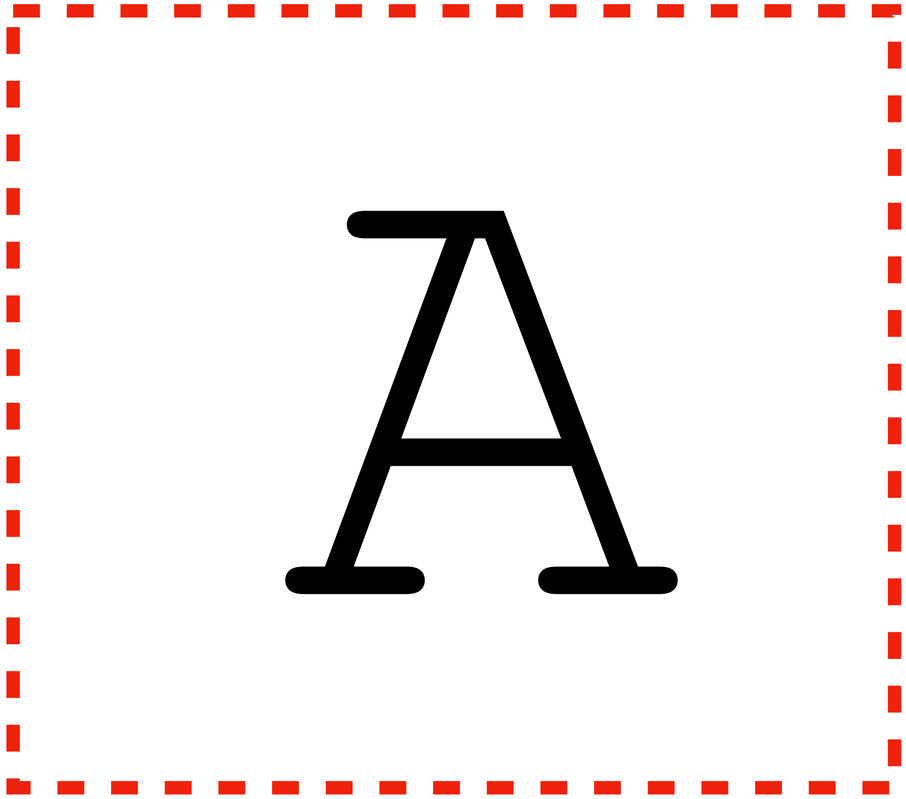
Resemblance to Known Brain Mechanisms



Traylor, Aaron, et al. "Transformer Mechanisms Mimic Frontostriatal Gating Operations When Trained on Human Working Memory Tasks." arXiv preprint arXiv:2402.08211 (2024).

Resemblance to Known Brain Mechanisms

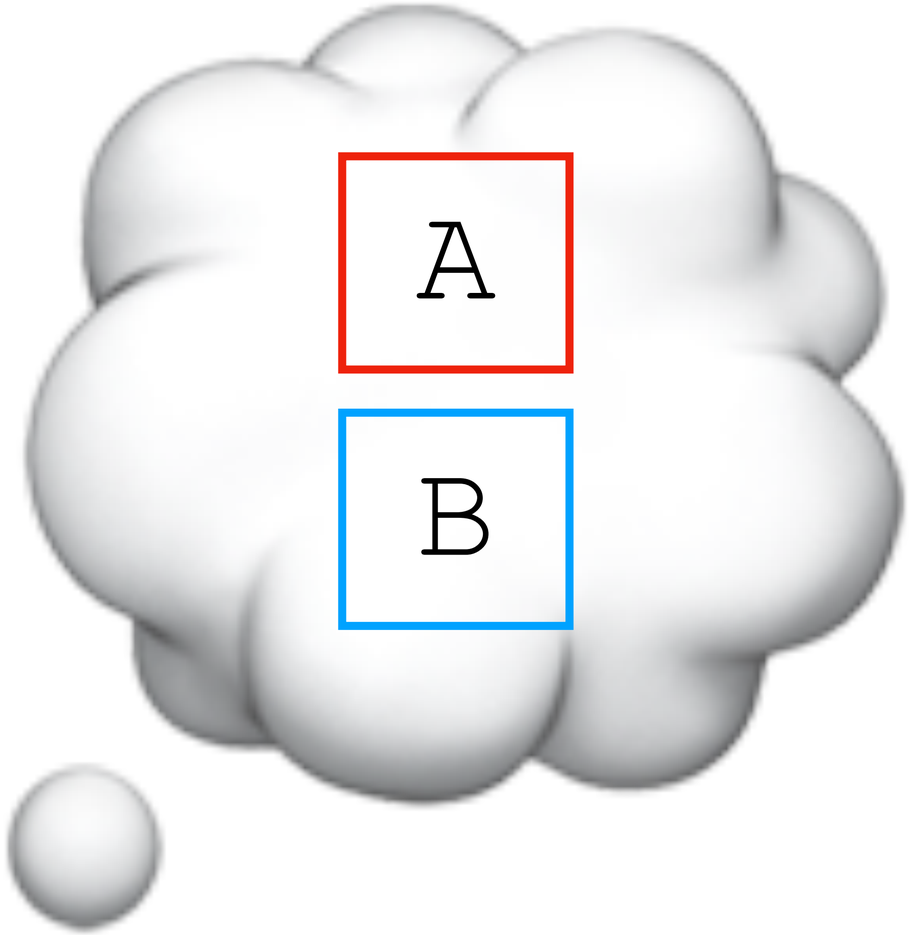


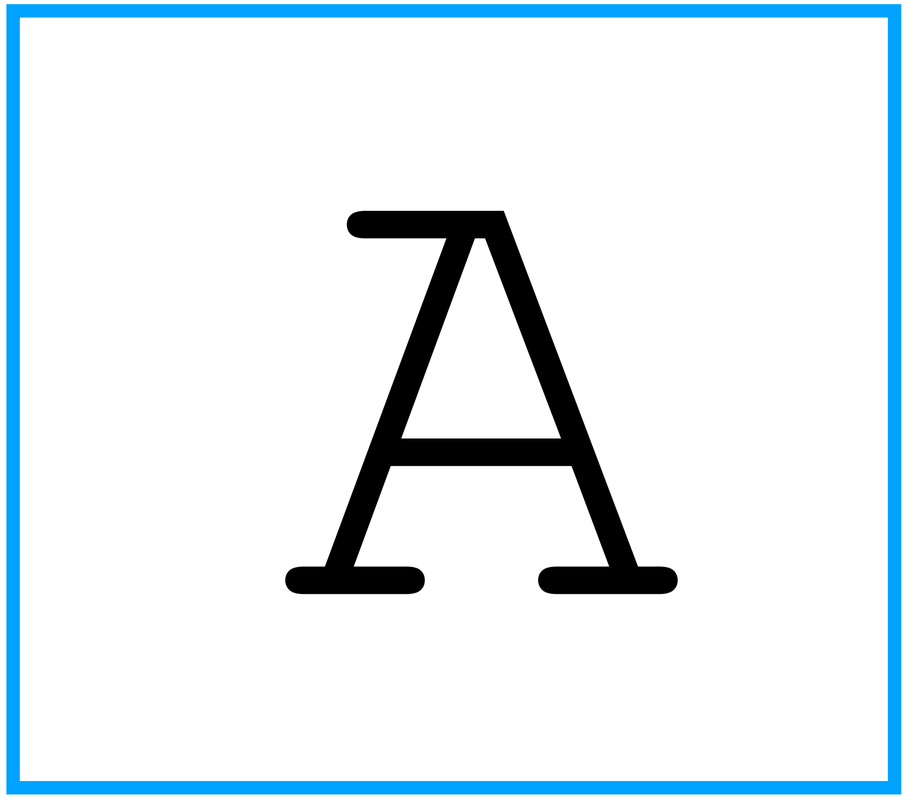


✅

Traylor, Aaron, et al. "Transformer Mechanisms Mimic Frontostriatal Gating Operations When Trained on Human Working Memory Tasks." arXiv preprint arXiv:2402.08211 (2024).

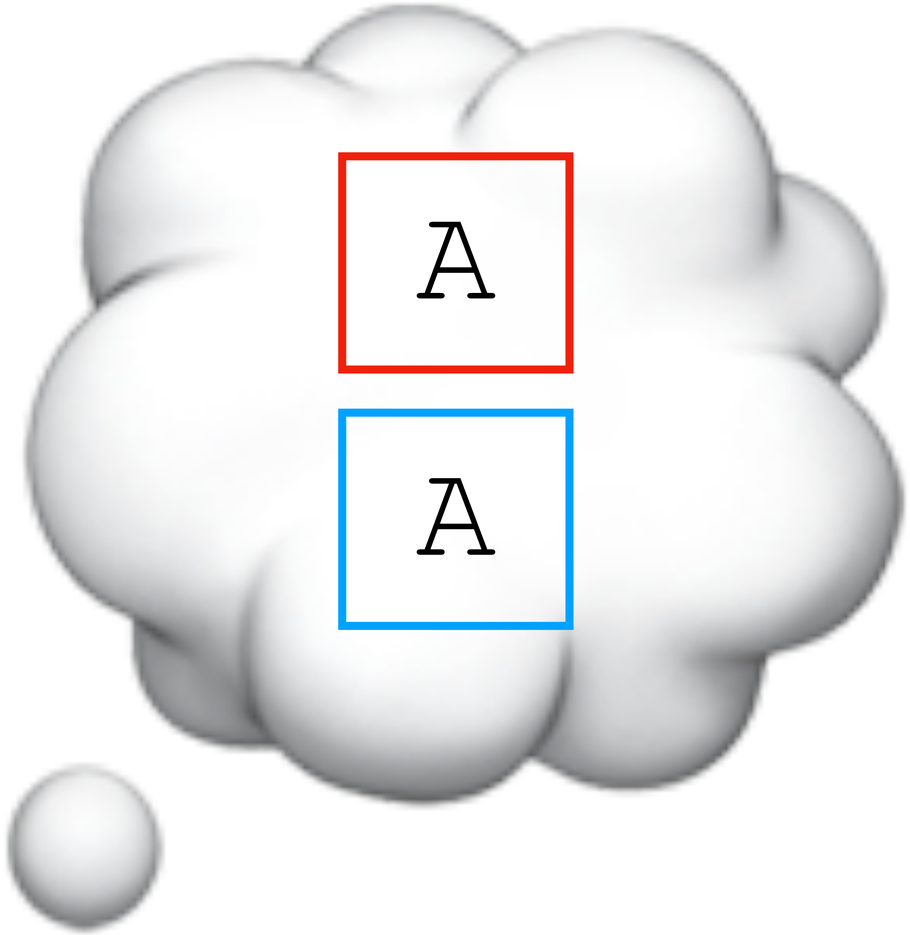
Resemblance to Known Brain Mechanisms

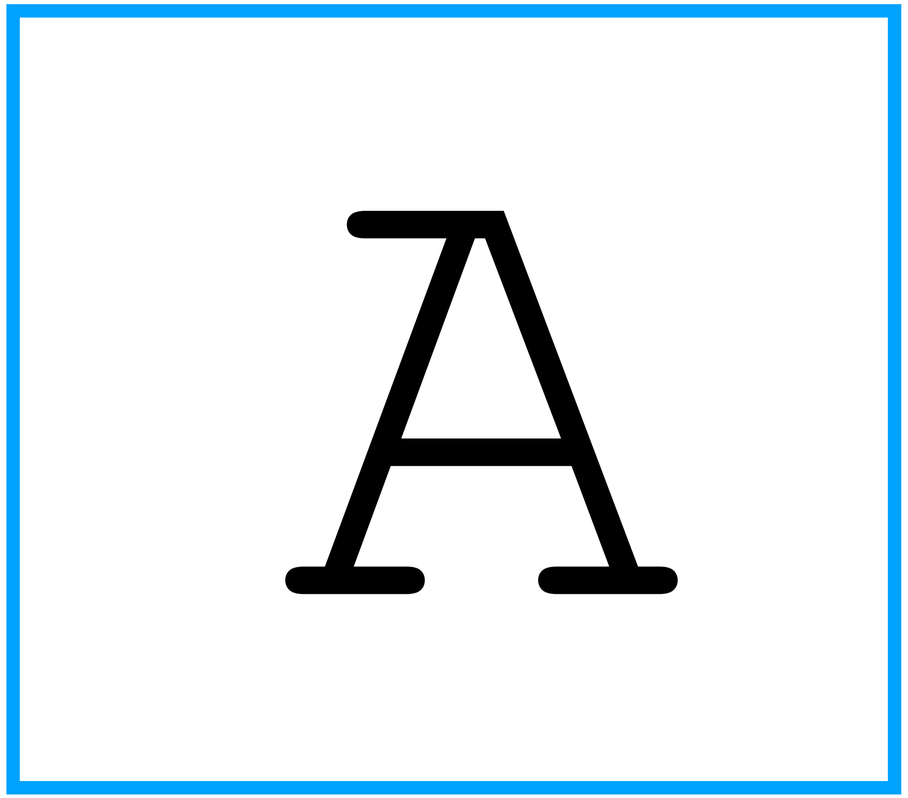




Traylor, Aaron, et al. "Transformer Mechanisms Mimic Frontostriatal Gating Operations When Trained on Human Working Memory Tasks." arXiv preprint arXiv:2402.08211 (2024).

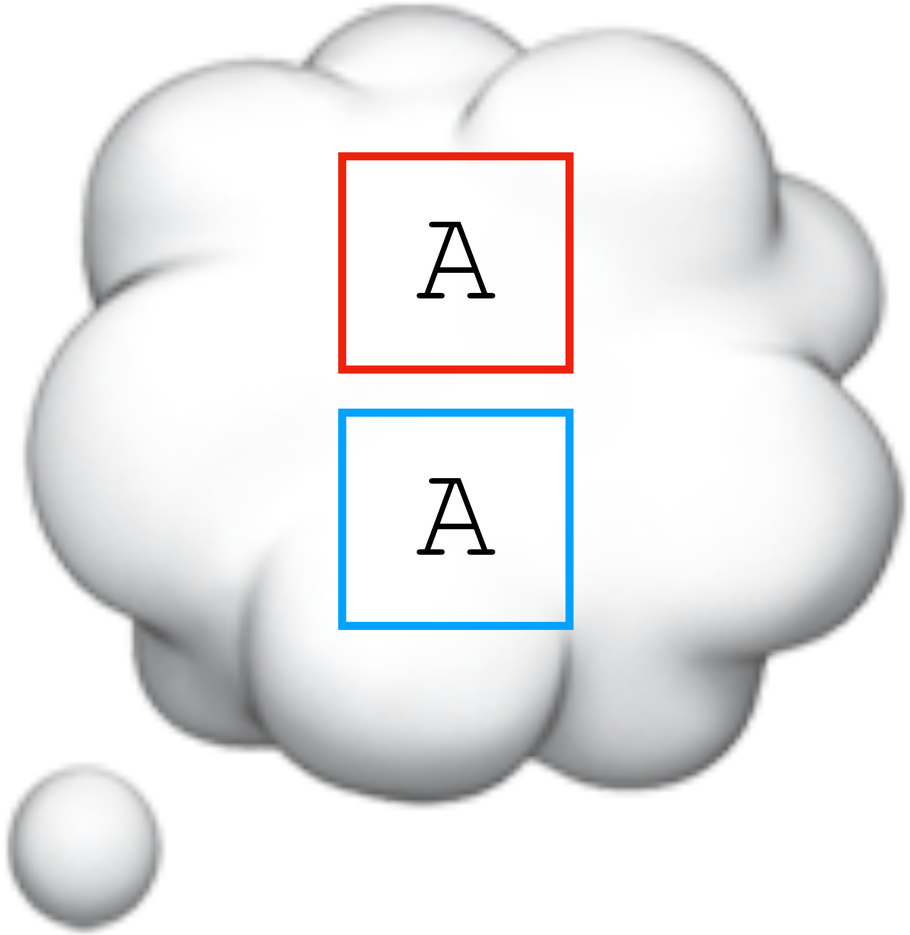
Resemblance to Known Brain Mechanisms





Traylor, Aaron, et al. "Transformer Mechanisms Mimic Frontostriatal Gating Operations When Trained on Human Working Memory Tasks." arXiv preprint arXiv:2402.08211 (2024).

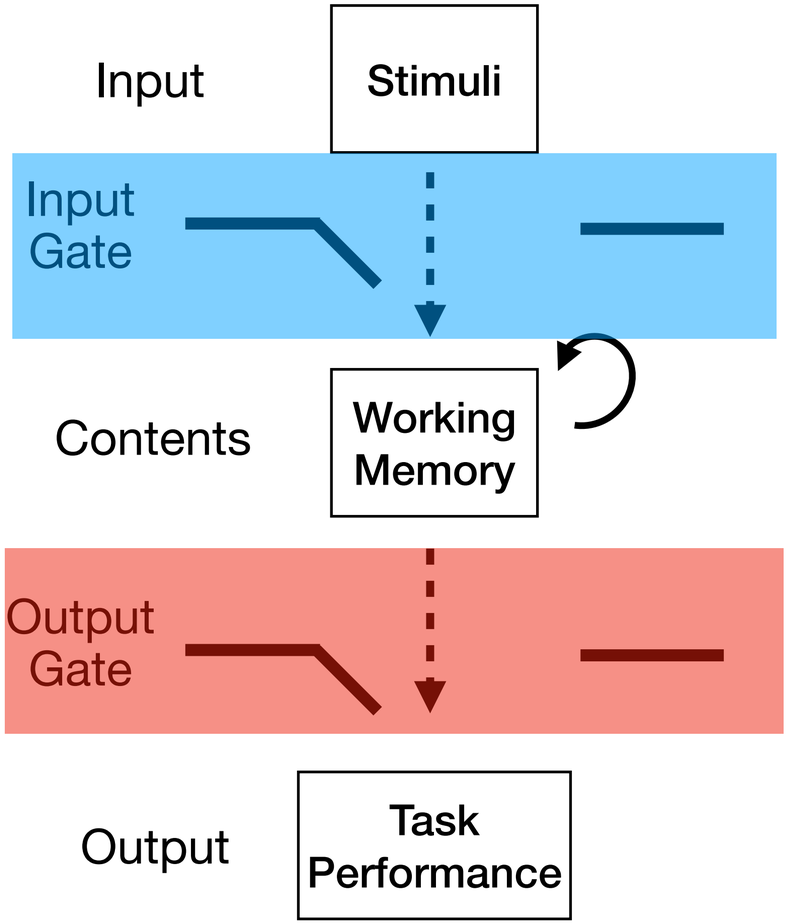
Resemblance to Known Brain Mechanisms



B

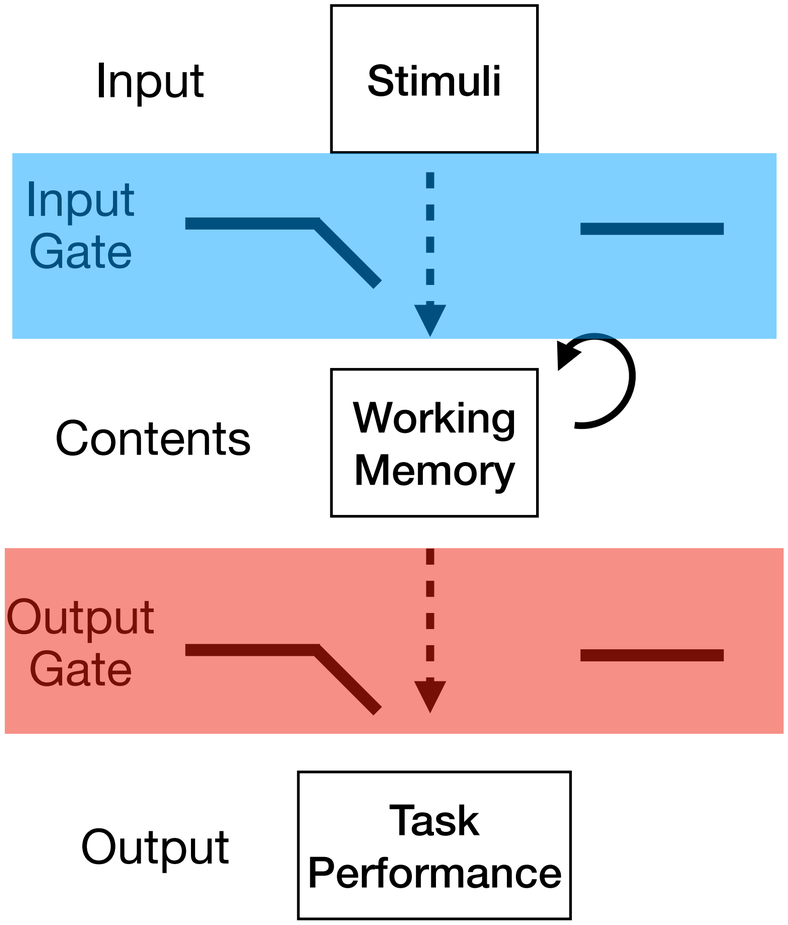
Traylor, Aaron, et al. "Transformer Mechanisms Mimic Frontostriatal Gating Operations When Trained on Human Working Memory Tasks." arXiv preprint arXiv:2402.08211 (2024).

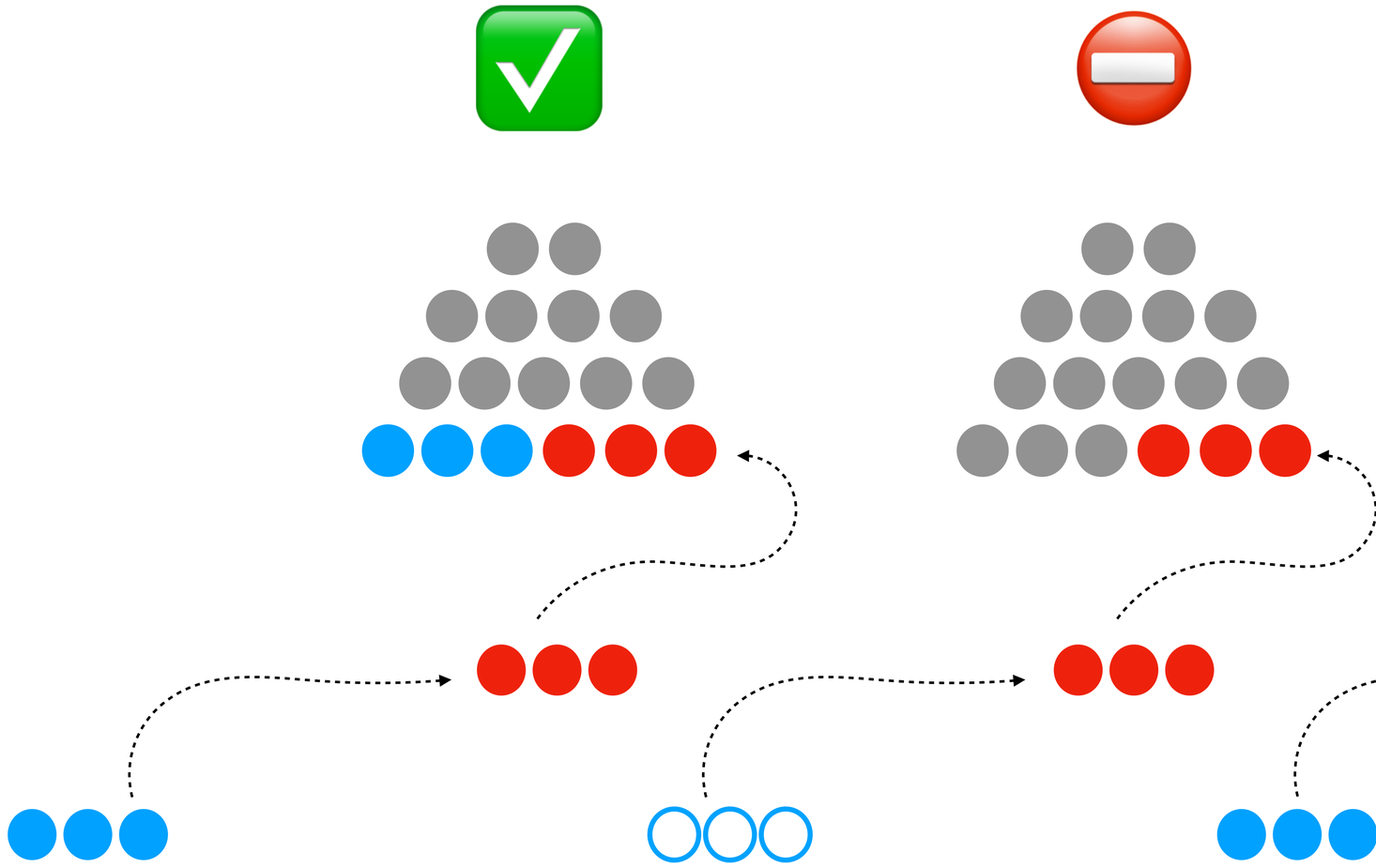
Resemblance to Known Brain Mechanisms



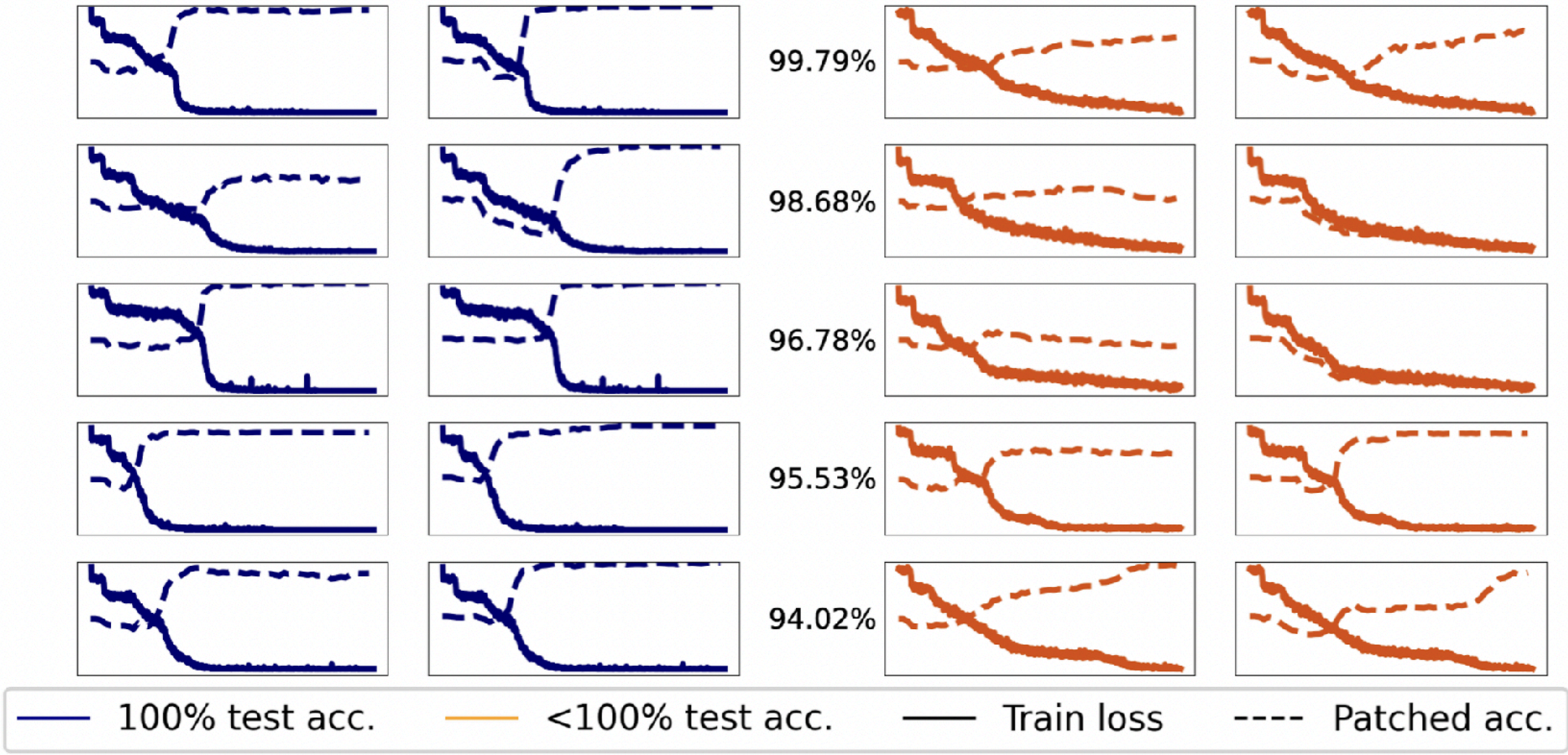
Traylor, Aaron, et al. "Transformer Mechanisms Mimic Frontostriatal Gating Operations When Trained on Human Working Memory Tasks." arXiv preprint arXiv:2402.08211 (2024).

Resemblance to Known Brain Mechanisms

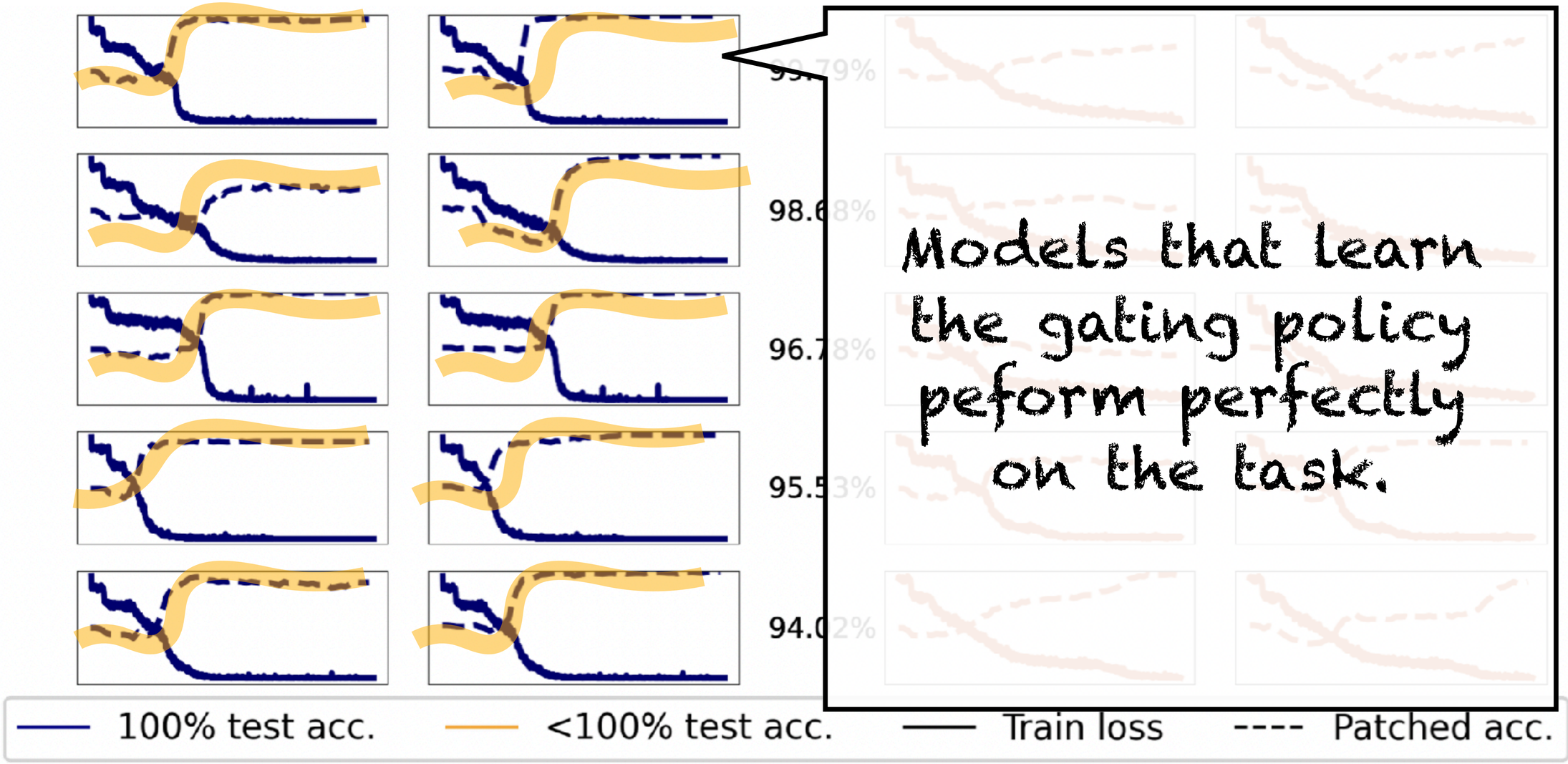




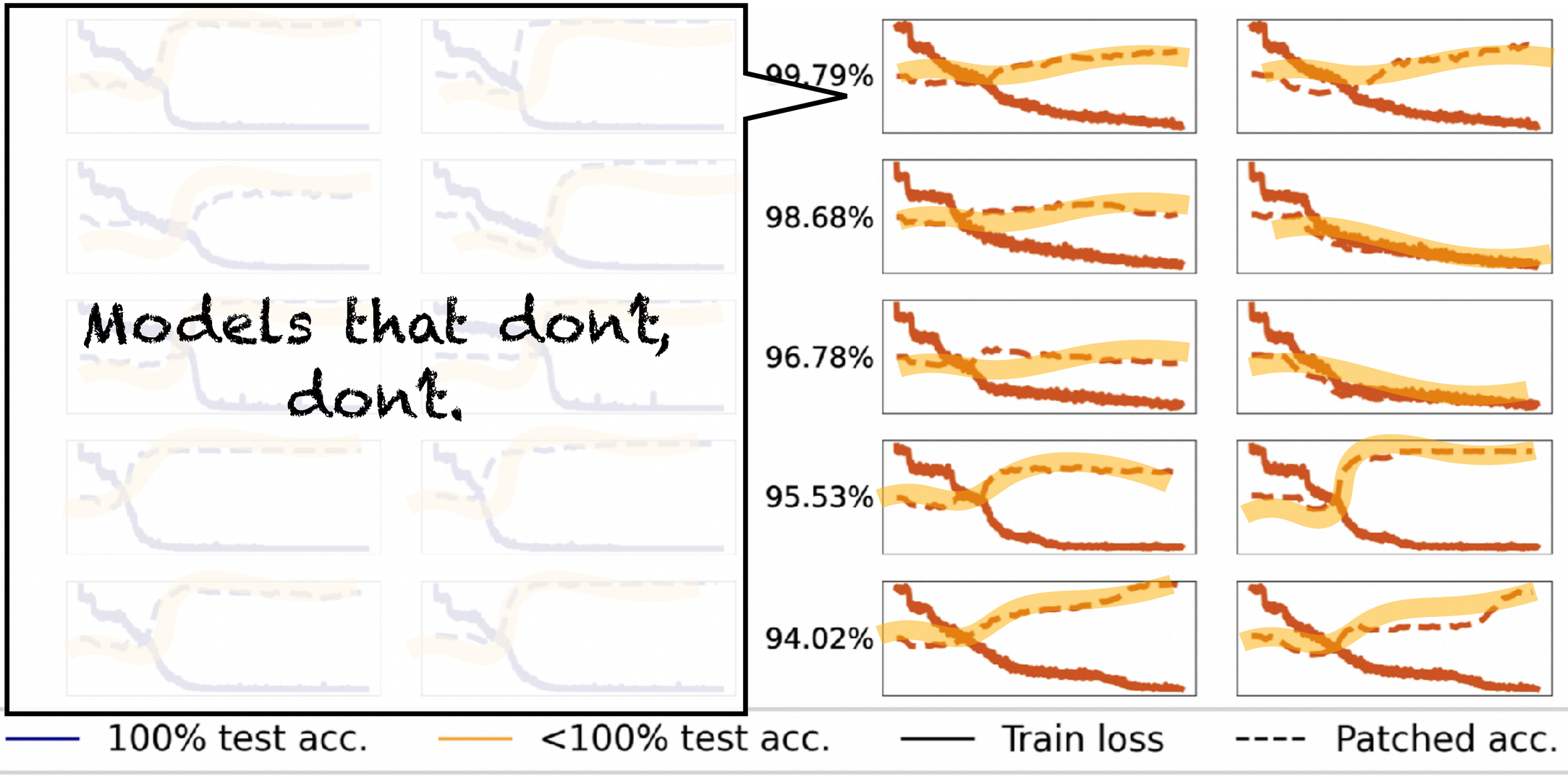
Resemblance to Known Brain Mechanisms



Resemblance to Known Brain Mechanisms



Resemblance to Known Brain Mechanisms



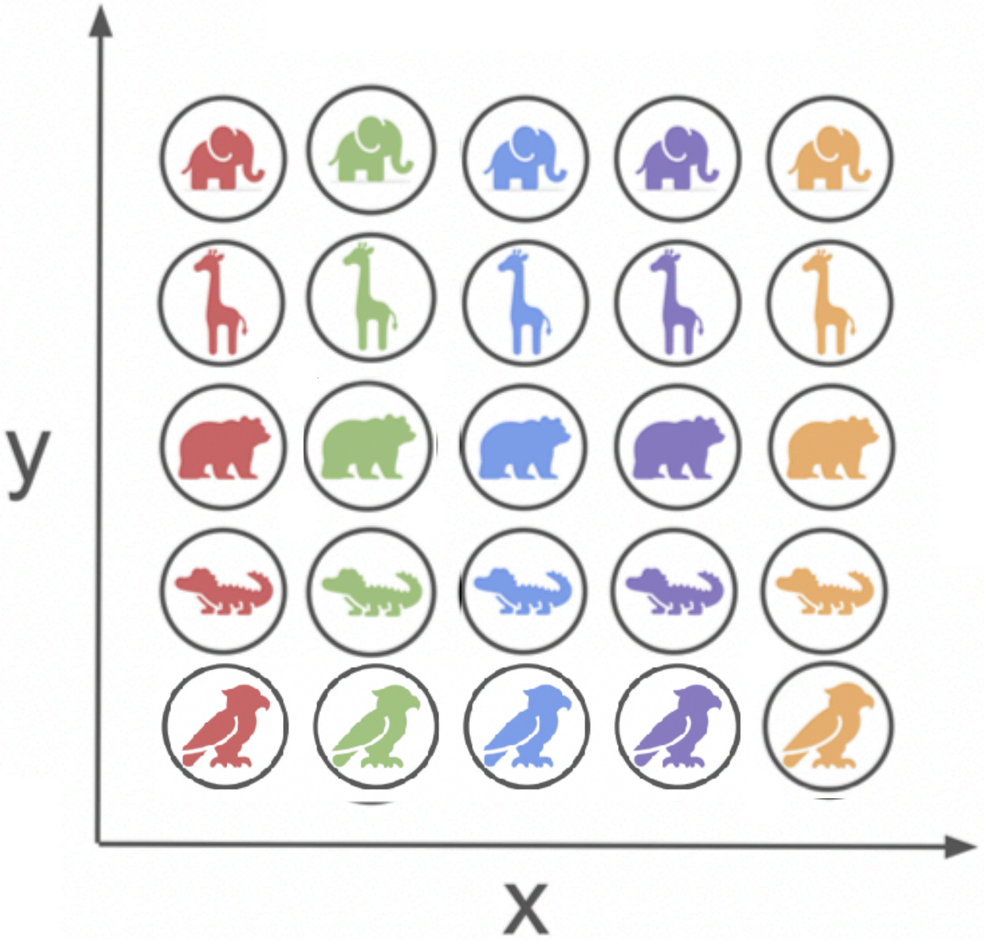
Proofs of Concept

- LLMs are not merely “stochastic parrots”. They often contain **interpretable internal structure** which makes use of modular functionally-specialized components.

- These components can **sometimes resemble known brain mechanisms**, e.g., mechanisms for cognitive control in the prefrontal cortex

- LLMs, off the shelf, exhibit **behaviors for which we don’t have good competing computational cognitive models**, like the ability to learn flexibly both in context and in weights. Understanding these mechanisms and aligning them to those of humans could jointly offer new theories of human cognition and improve LLMs robustness, interpretability, and efficiency

Ability to Explain New Behavioral Data



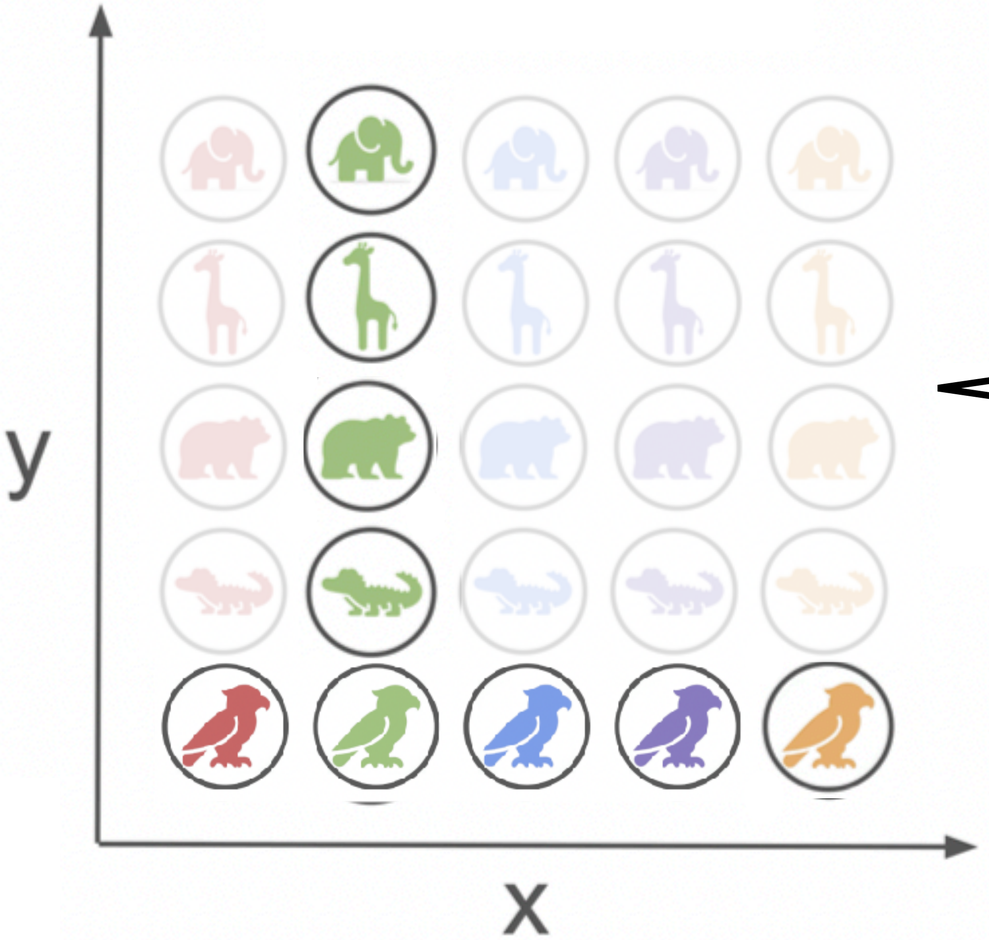

```
blue elephant : 3 5
blue giraffe : 3 4
blue bird : 3 1
blue alligator : 3 2
blue bear : 3 3
green bear : 2 3
purple bear : 4 3
orange bear : 5 3
red bear : 1 3
orange elephant :
```

study examples

query

Dekker, Ronald B., Fabian Otto, and Christopher Summerfield. "Curriculum learning for human compositional generalization." Proceedings of the National Academy of Sciences 119.41 (2022): e2205582119.

Ability to Explain New Behavioral Data



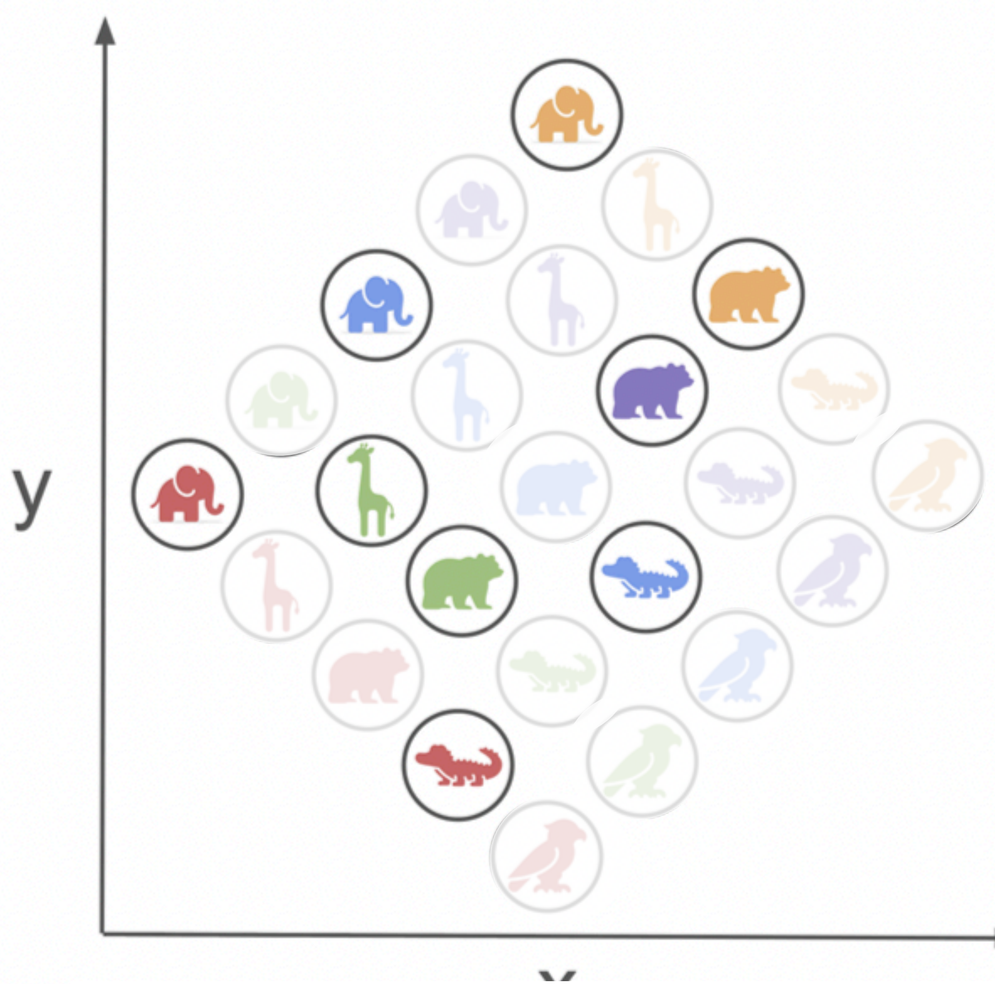
When task follows a "rule like" structure, humans prefer to learn from "blocked" examples.

study examples

query

Bokker, Ronald B., Fabian Otto, and Christopher Summerfield. "Curriculum learning for human compositional generalization." Proceedings of the National Academy of Sciences 119.41 (2022): e2205582119.

Ability to Explain New Behavioral Data



When task doesn't follow simple "rules", humans prefer shuffled examples.

blue elephant : 3 5
blue giraffe : 3 4
blue bird : 3 1
blue alligator : 3 2
green bear : 2 3
purple bear : 4 3
orange bear : 5 3
red bear : 1 3
orange elephant :

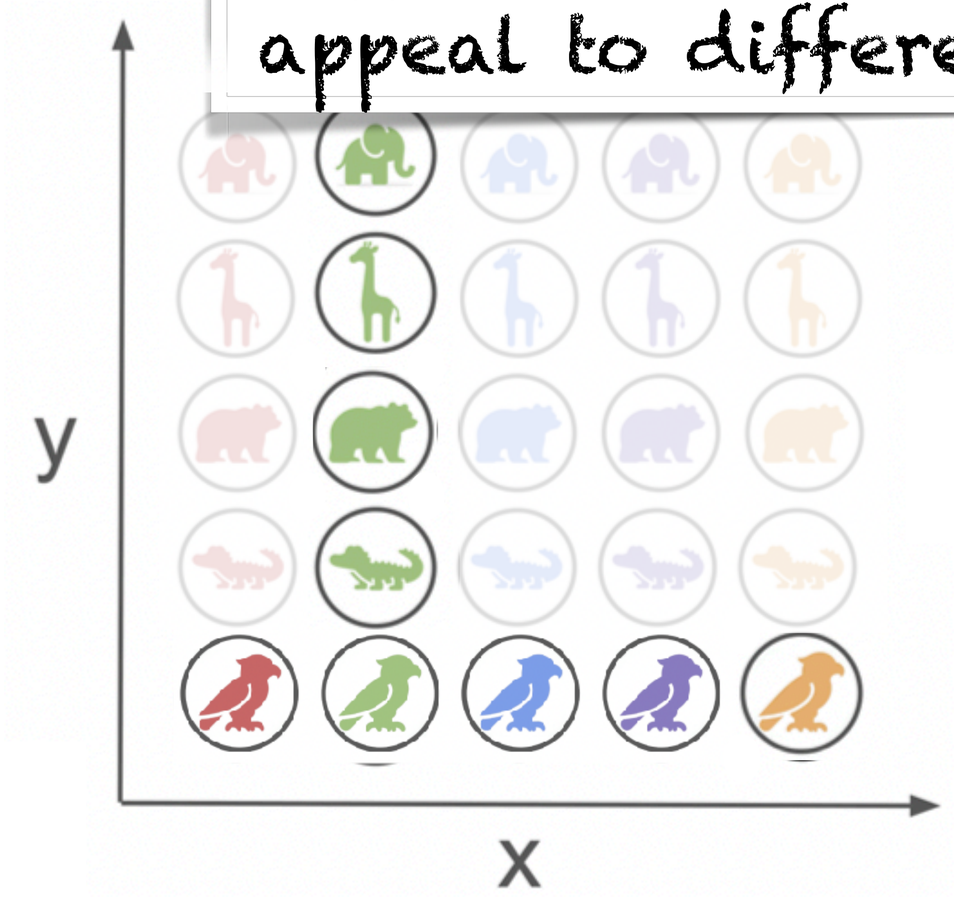
study examples

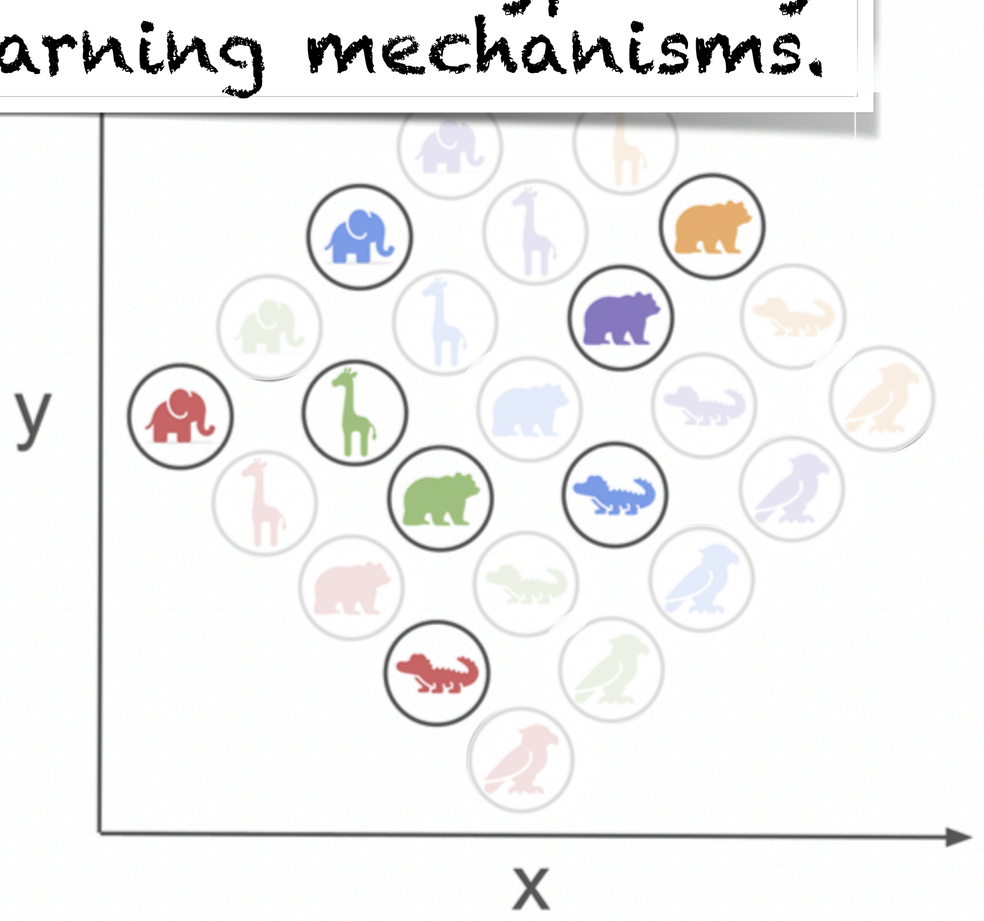
query

Noh, Sharon M., et al. "Optimal sequencing during category learning: Testing a dual-learning systems perspective." Cognition 155 (2016): 23-29.

Ability to Explain New Behavioral Data

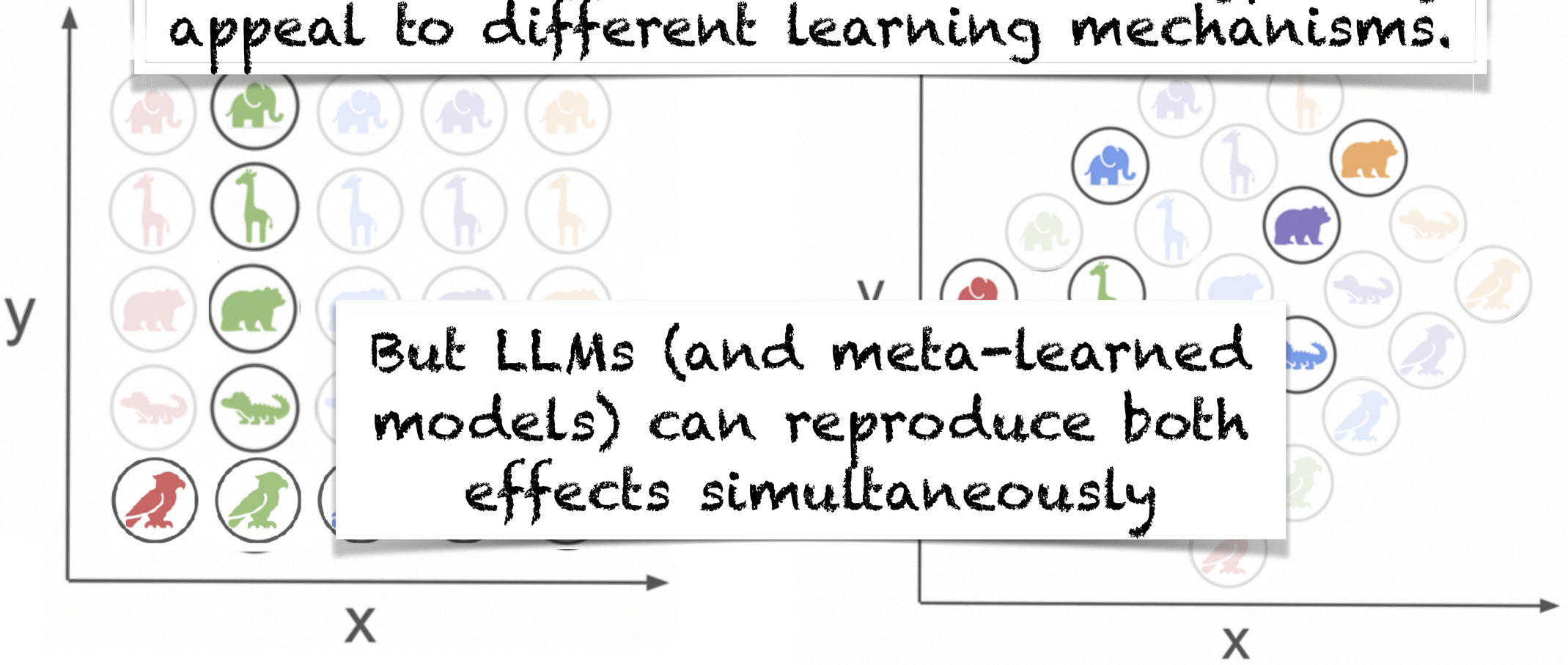
Different learning effects which typically appeal to different learning mechanisms.





Ability to Explain New Behavioral Data

Different learning effects which typically appeal to different learning mechanisms.



But LLMs (and meta-learned models) can reproduce both effects simultaneously

Russin, Jacob, Ellie Pavlick, and Michael J. Frank. "Human Curriculum Effects Emerge with In-Context Learning in Neural Networks." arXiv preprint arXiv:2402.08674 (2024).

Proofs of Concept

- LLMs are not merely “stochastic parrots”. They often contain **interpretable internal structure** which makes use of modular functionally-specialized components.
- These components can **sometimes resemble known brain mechanisms**, e.g., mechanisms for cognitive control in the prefrontal cortex
- LLMs, off the shelf, exhibit **behaviors for which we don’t have good competing computational cognitive models**, like the ability to learn flexibly both in context and in weights. Understanding these mechanisms and aligning them to those of humans could jointly offer new theories of human cognition and improve LLMs robustness, interpretability, and efficiency

Moving Forward

- Using LLMs as cognitive models has the potential to unlock a **virtuous cycle** that both improves our understanding of human cognition and improves AI
- Harnessing this requires:
  - **Open source models** for research which match the scale and performance of proprietary models
  - Investment in research with a **long time horizon**, which falls outside of commercial priorities (decades, not years)
  - (Even more) open-minded, **interdisciplinary** collaboration

Thank you!