Toward an Understanding of Symbol Processing in Neural Network Architectures

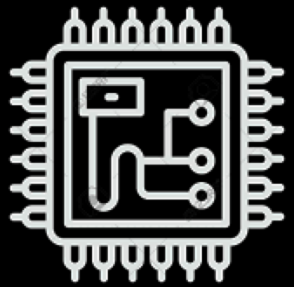
Jonathan D. Cohen

Princeton Neuroscience Institute
Princeton AI Lab

Representing:

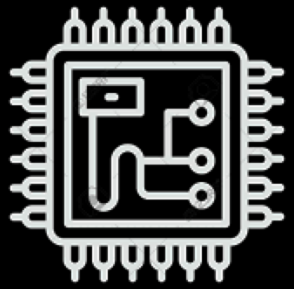
Awni Altabaa *(Computer Science, Yale)*
Declan Campbell *(Neuroscience, Princeton)*
Steven Frankland *(Cognitive Science, Dartmouth)*
Tyler Giallanza *(Psychology, Princeton)*
Kamesh Krishnamurthy *(Physics and Neuroscience, Princeton)*
John Lafferty *(Computer Science, Yale)*
Shanka Mondal *(Electrical Engineering, Princeton)*
Simon Segert *(Neuroscience, Princeton)*
Taylor Webb *(Psychology, Microsoft Research)*
Yukang Yan *(Electrical Engineering, Princeton)*

Current State of Affairs



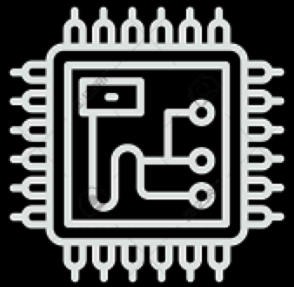
Current State of Affairs

- **Miracle of symbolic computing:**
  - 👍 **Computationally general: maximum flexibility**
    Interpretable, reliable



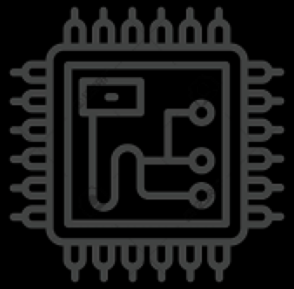
Current State of Affairs

- **Miracle of symbolic computing:**
  - 👍 Computationally general: maximum flexibility
    Interpretable, reliable
  - 👎 Inefficient / difficult to configure for hyper-complex domains
    – vision, natural language

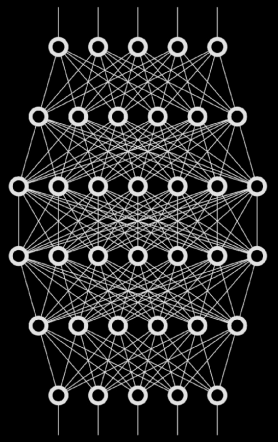


Current State of Affairs

- **Miracle of symbolic computing:**
  - 👍 Computationally general: maximum flexibility
    Interpretable, reliable
  - 👎 Inefficient / difficult to configure for hyper-complex domains
    – vision, natural language

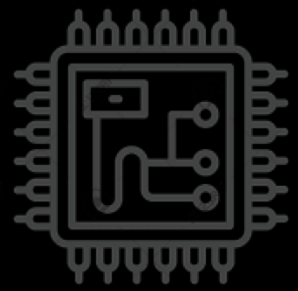


- **Miracle of deep learning:**
  - 👍 Computationally efficient: automated function approximation



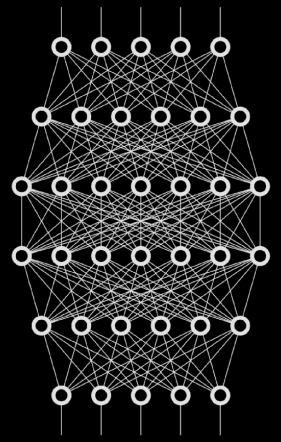
Current State of Affairs

- **Miracle of symbolic computing:**



  - 👍 Computationally general: maximum flexibility
    Interpretable, reliable
  - 👎 Inefficient / difficult to configure for hyper-complex domains
    – vision, natural language

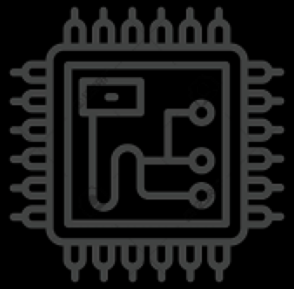
- **Miracle of deep learning:**



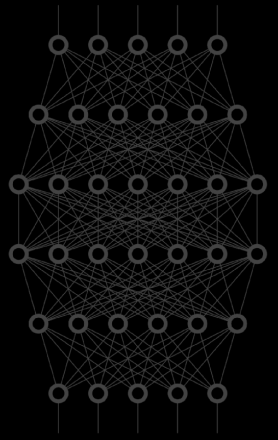
  - 👍 Computationally efficient: automated function approximation
  - 👎 Sample-inefficient
    Domain-specific
    Interpretable? reliable?

Current State of Affairs

- **Miracle of symbolic computing:**
  - 👍 Computationally general: maximum flexibility
    Interpretable, reliable
  - 👎 Inefficient / difficult to configure for hyper-complex domains
    – vision, natural language



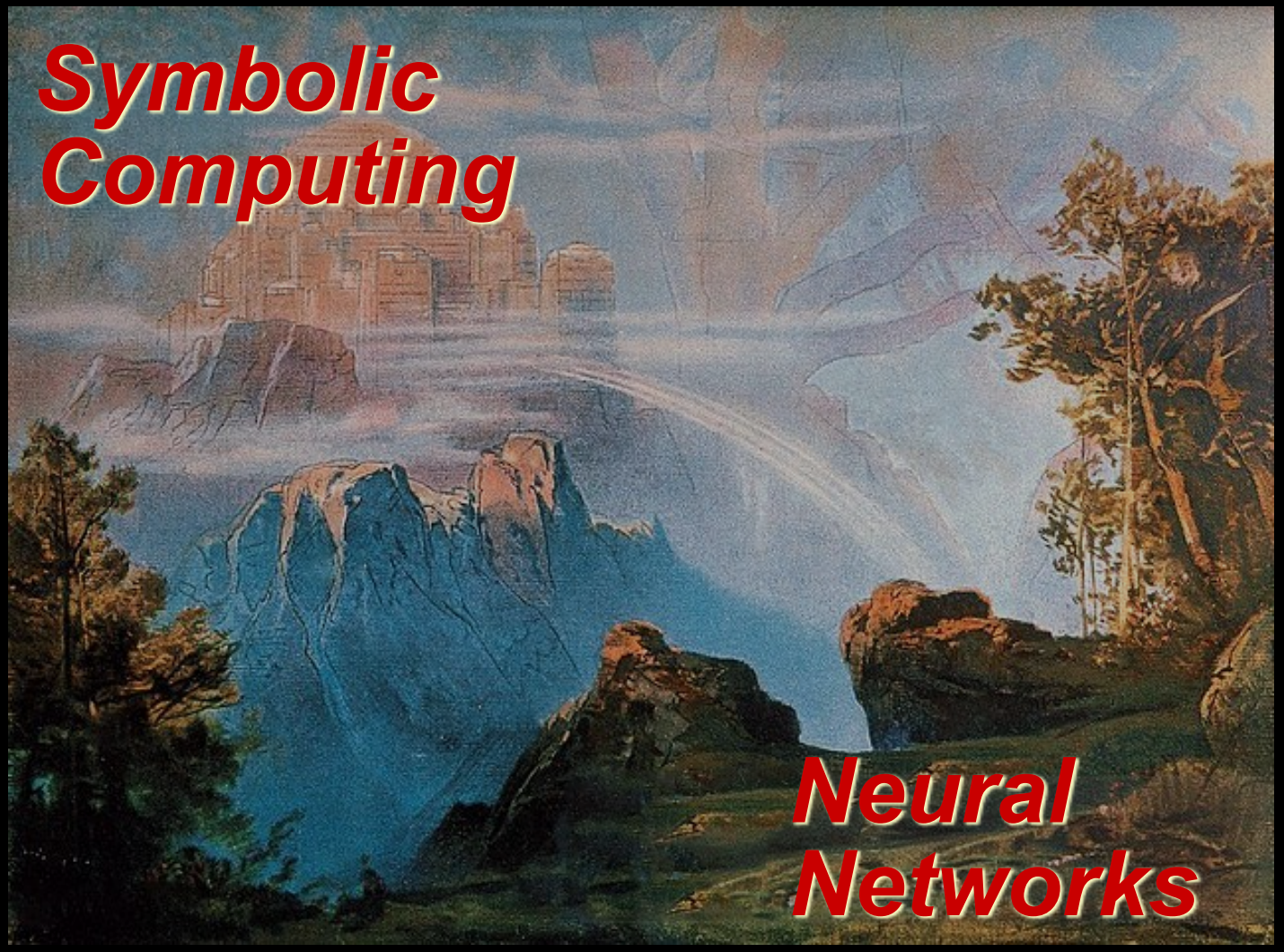
- **Miracle of deep learning:**
  - 👍 Computationally efficient: automated function approximation
  - 👎 Sample-inefficient
    Domain-specific
    Interpretable? reliable?



- **So where are we?**

Clash of the Titans

Shangri-La




Shangri-La?

Shangri-La?

- **Challenge:**
  - Integrate ***flexibility*** of *symbolic processing* in traditional architectures
  - with ***efficiency*** of *function approximation* in neural networks

Shangri-La?

- **Current efforts:**
  - **Neuro-symbolic approaches:**
    - start with pre-specified **symbolic primitives** (“core knowledge”)
    - use deep learning to **combine these** (e.g., “program induction”)

Shangri-La?

- **Current efforts:**
  - **"Neo-connectionist" approaches:**
    - use deep learning for "end-to-end" training of neural networks

Shangri-La?

- **Current efforts:**

  - **"Neo-connectionist" approaches:**

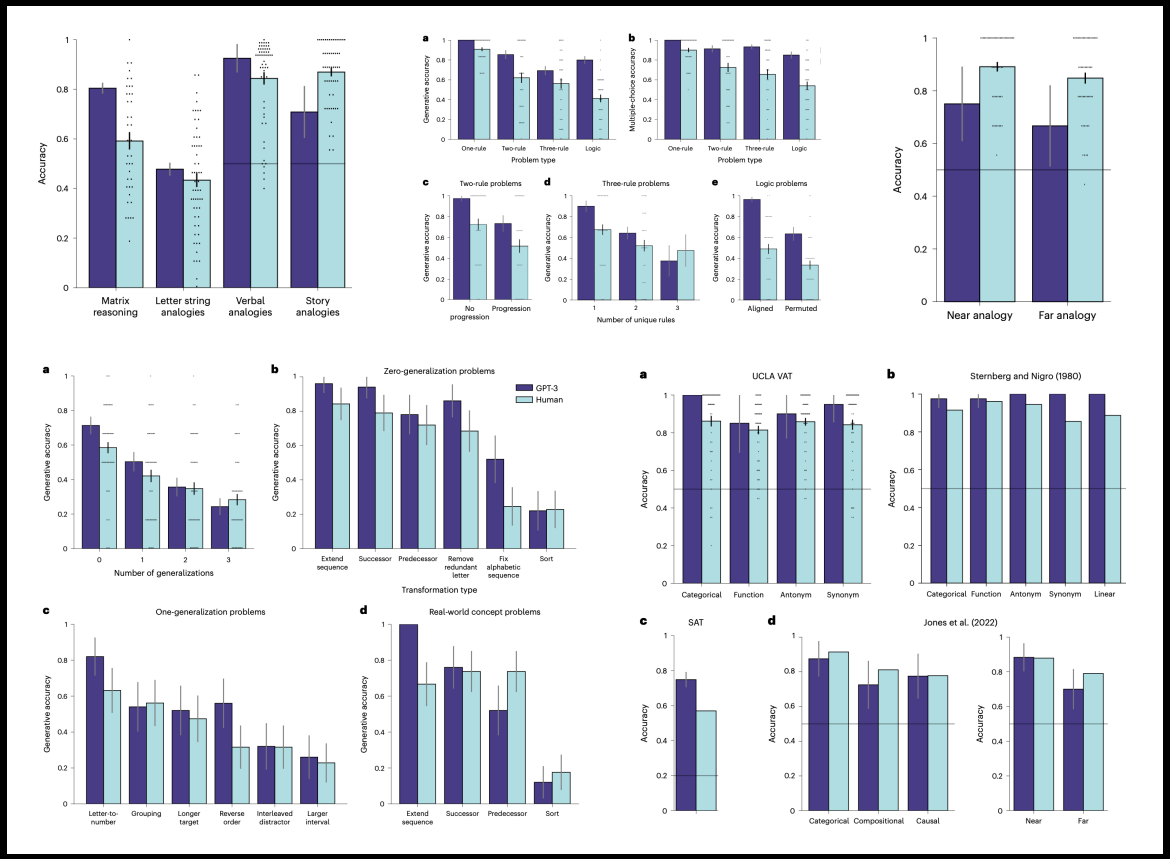
    - inductive biases that favor **abstraction**
      - *training:* curricular learning, meta learning
      - *architecture & processing:* attention, external memory

Shangri-La?

- What about large neural networks?

Growing Evidence of Abstract Reasoning in NNs

(Webb et al, 2024)



GPT-3

Human

Growing Evidence of Abstract Reasoning in NNs

(Webb et al, 2024)

- **But wildly data- (and energy) inefficient**

Growing Evidence of Abstract Reasoning in NNs

(Webb et al, 2024)

- But wildly data- (and energy) inefficient
  and largely domain-specific

Growing Evidence of Abstract Reasoning in NNs

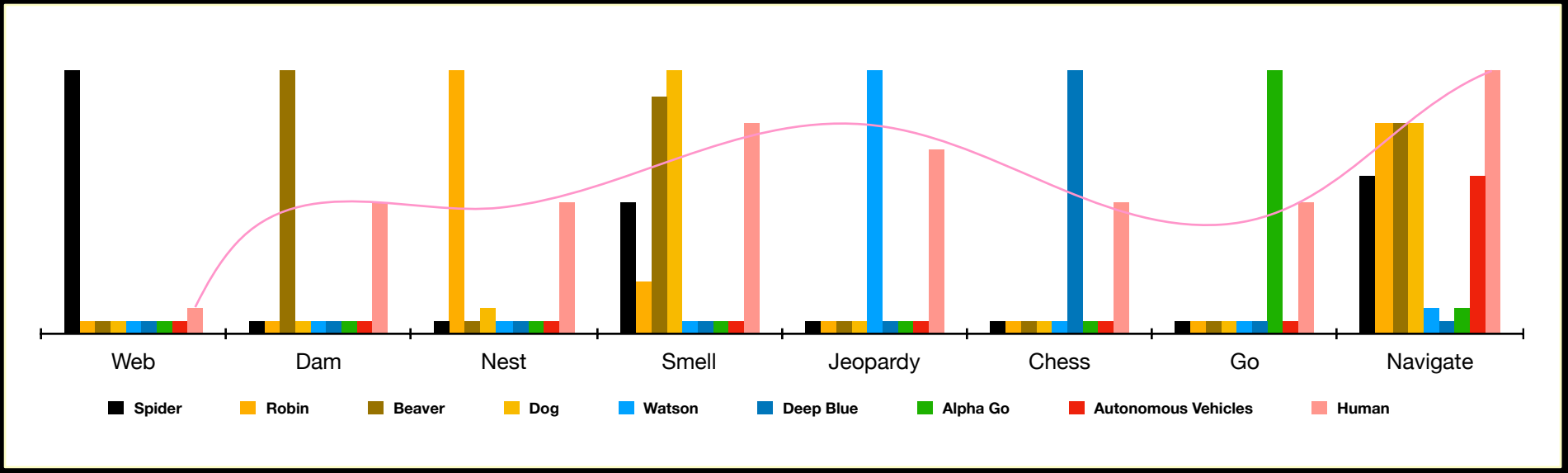
(Webb et al, 2024)

- But wildly data- (and energy) inefficient and largely domain-specific

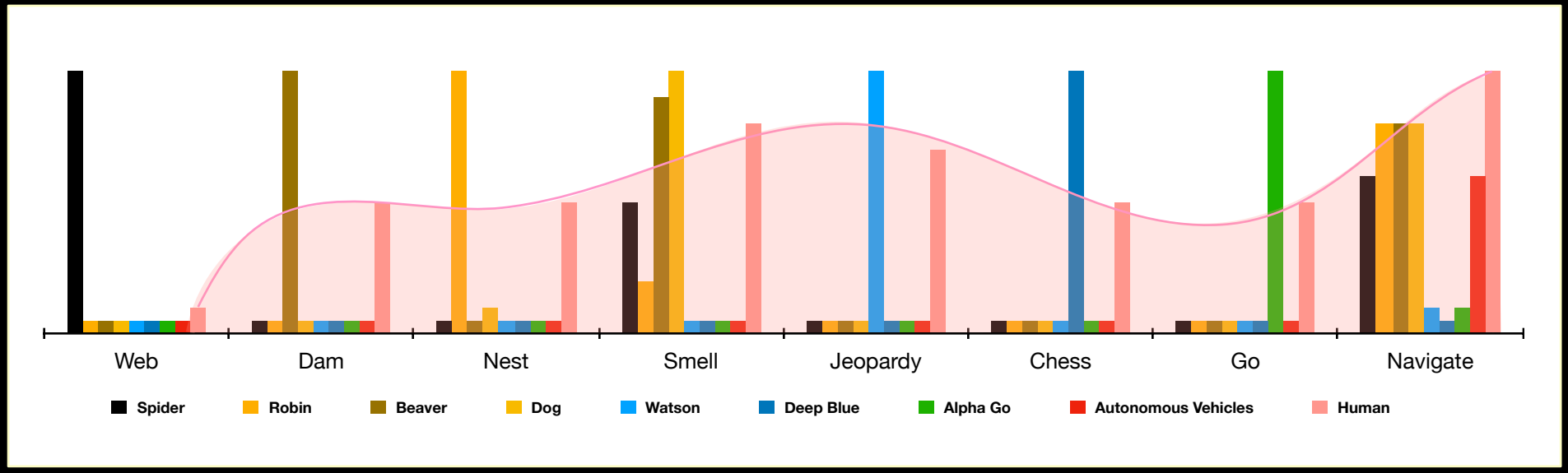
but there is an existence proof…

The Human Brain

The Human Brain



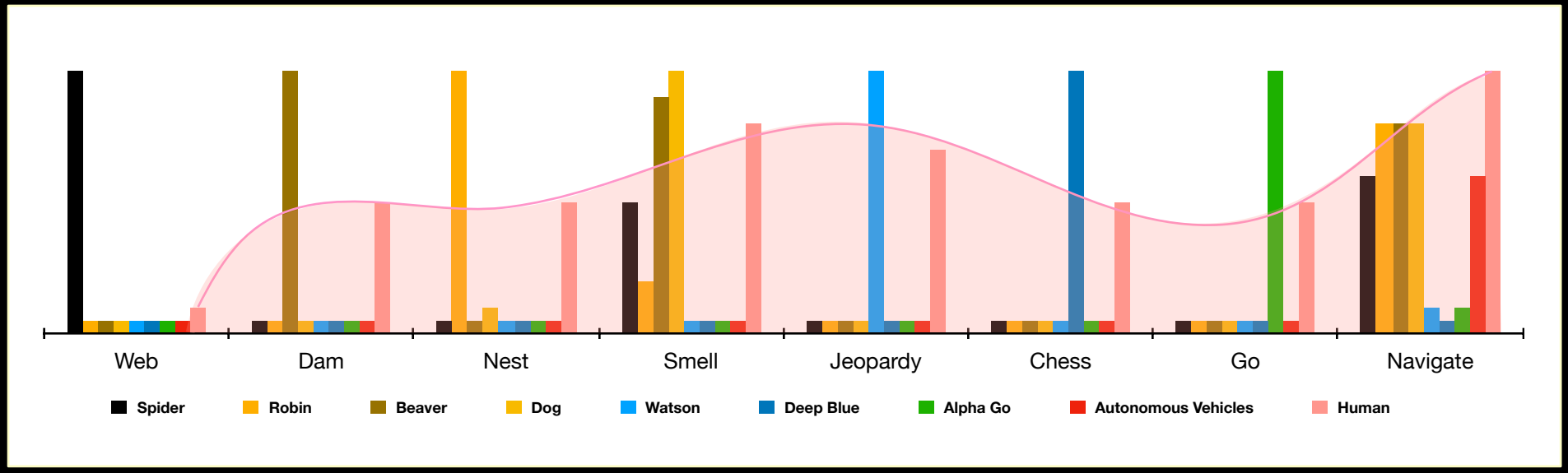
The Human Brain



"*Sweet spot*" between flexibility and efficiency

- Near limitless range of tasks at adequate performance - flexibility

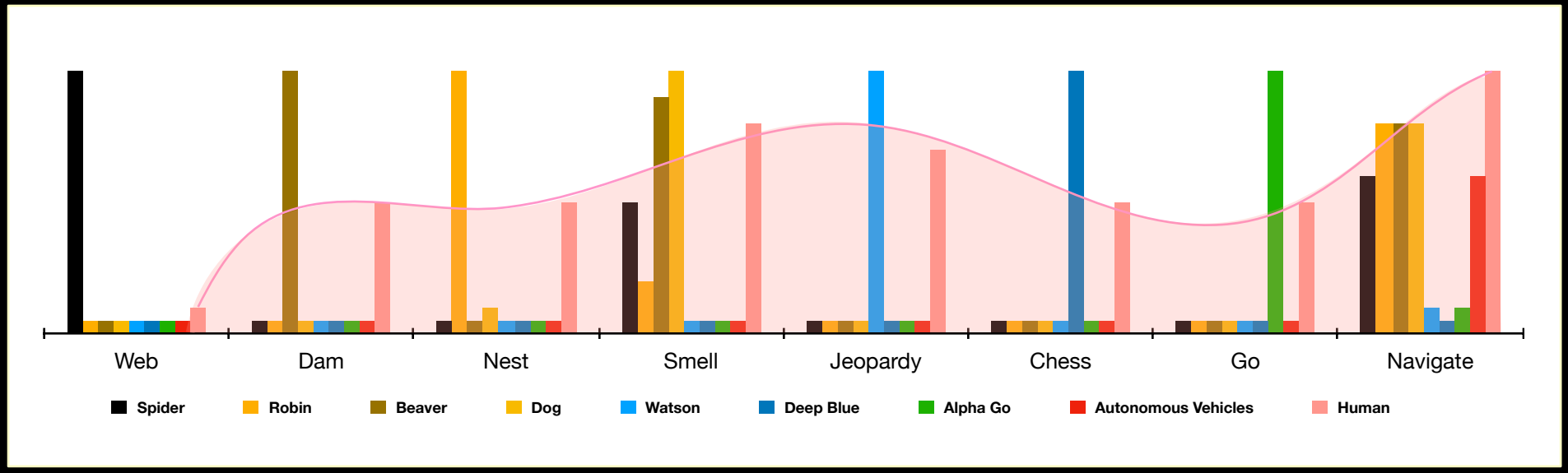
The Human Brain



"*Sweet spot*" between flexibility and efficiency

- With *reasonable* amounts, and often little or *no* training - sample efficiency

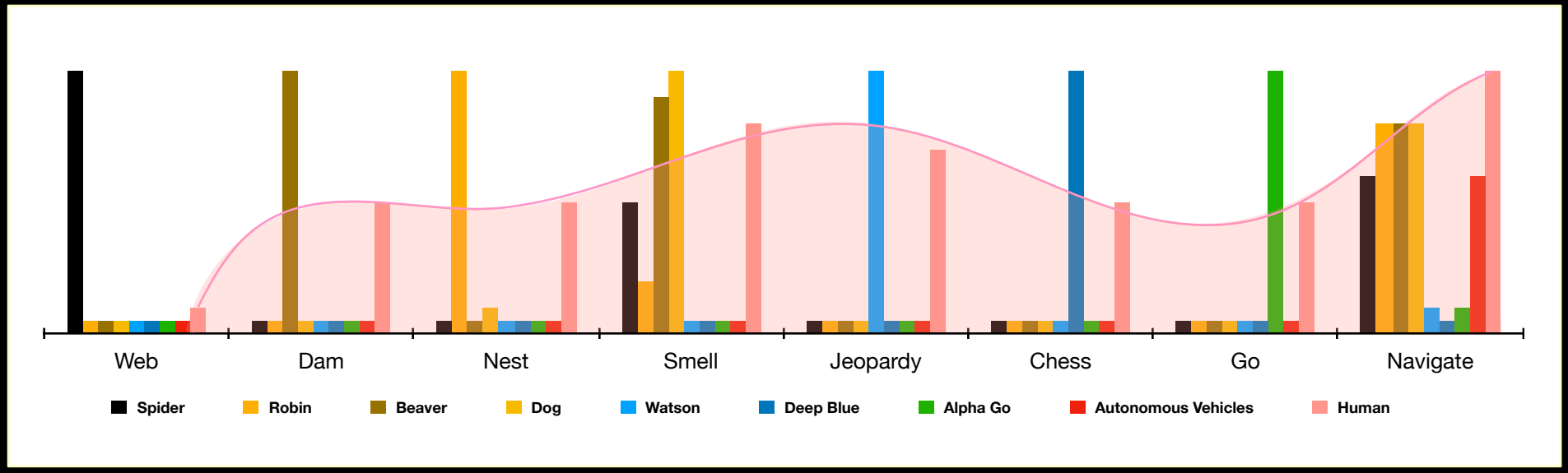
The Human Brain



"*Sweet spot*" between flexibility and efficiency

- ~20 watts, often with parallel performance - processing efficiency

The Human Brain



How does it accomplish this?

Outline

Outline

- Propose an architectural principle

Outline

- Propose an architectural principle
- Show that it works

Outline

- Propose an architectural principle
- Show that it works
- **Show how it works**

Outline

- Propose an architectural principle
- Show that it works
- Show how it works
- **Discuss:**
  - Generalizability

Outline

- Propose an architectural principle
- Show that it works
- Show how it works
- **Discuss:**
  - Generalizability
  - Relationship to symbolic (production system) architectures

Outline

- Propose an architectural principle
- Show that it works
- Show how it works
- Discuss:
  - Generalizability
  - Relationship to symbolic (production system) architectures
  - Biological plausibility (for those who care ;^)

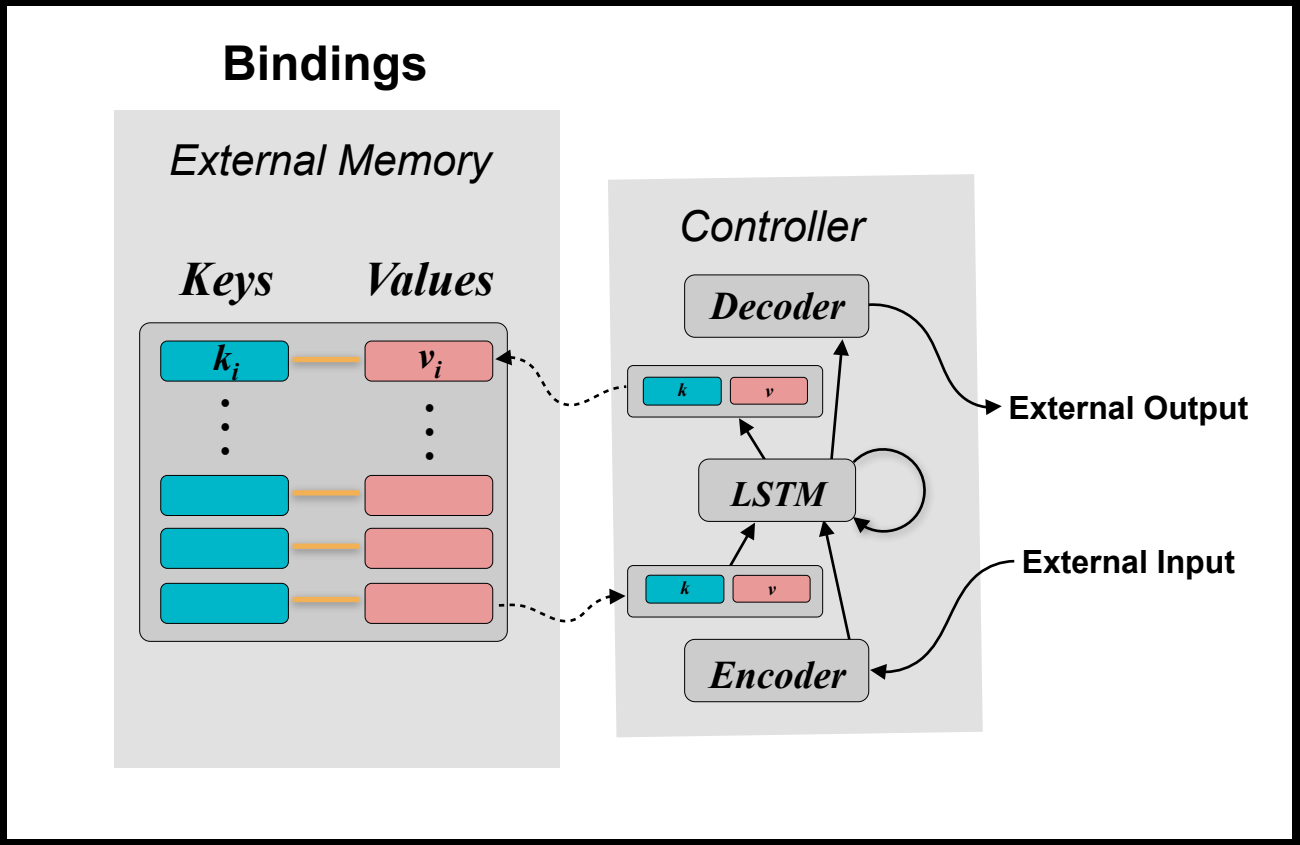
Outline

- **Propose an architectural principle**
- **Show that it works**
- **Show how it works**
- **Discuss:**
  - Generalizability
  - Relationship to symbolic (production system) architectures
  - Biological plausibility (for those who care ;^)
  - Variations on the theme (if there's time)

Neural Network with External Memory

Neural Turning Machine (NTN)

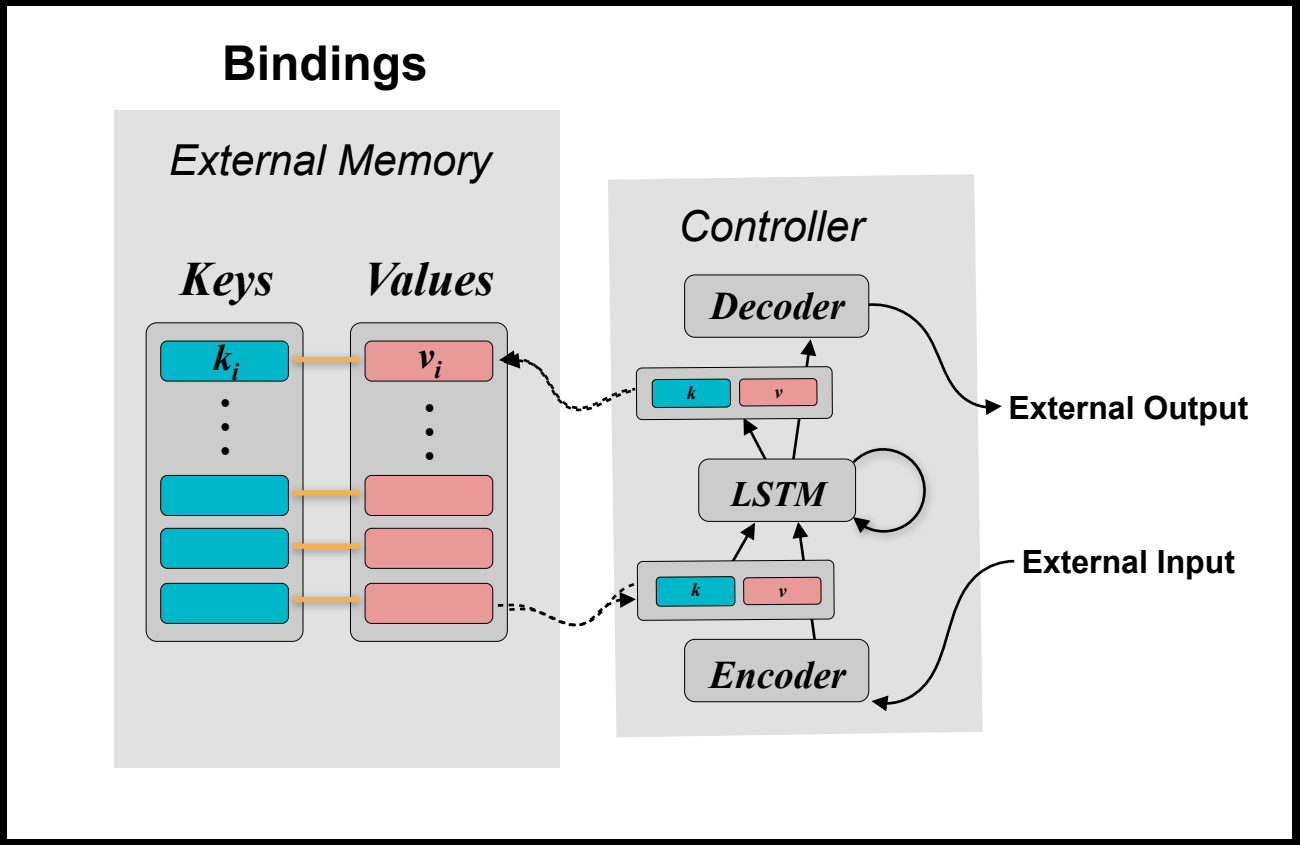
(Graves et al., 2014)



Neural Network with External Memory

Emergent Symbols Through Binding Network (ESBN)

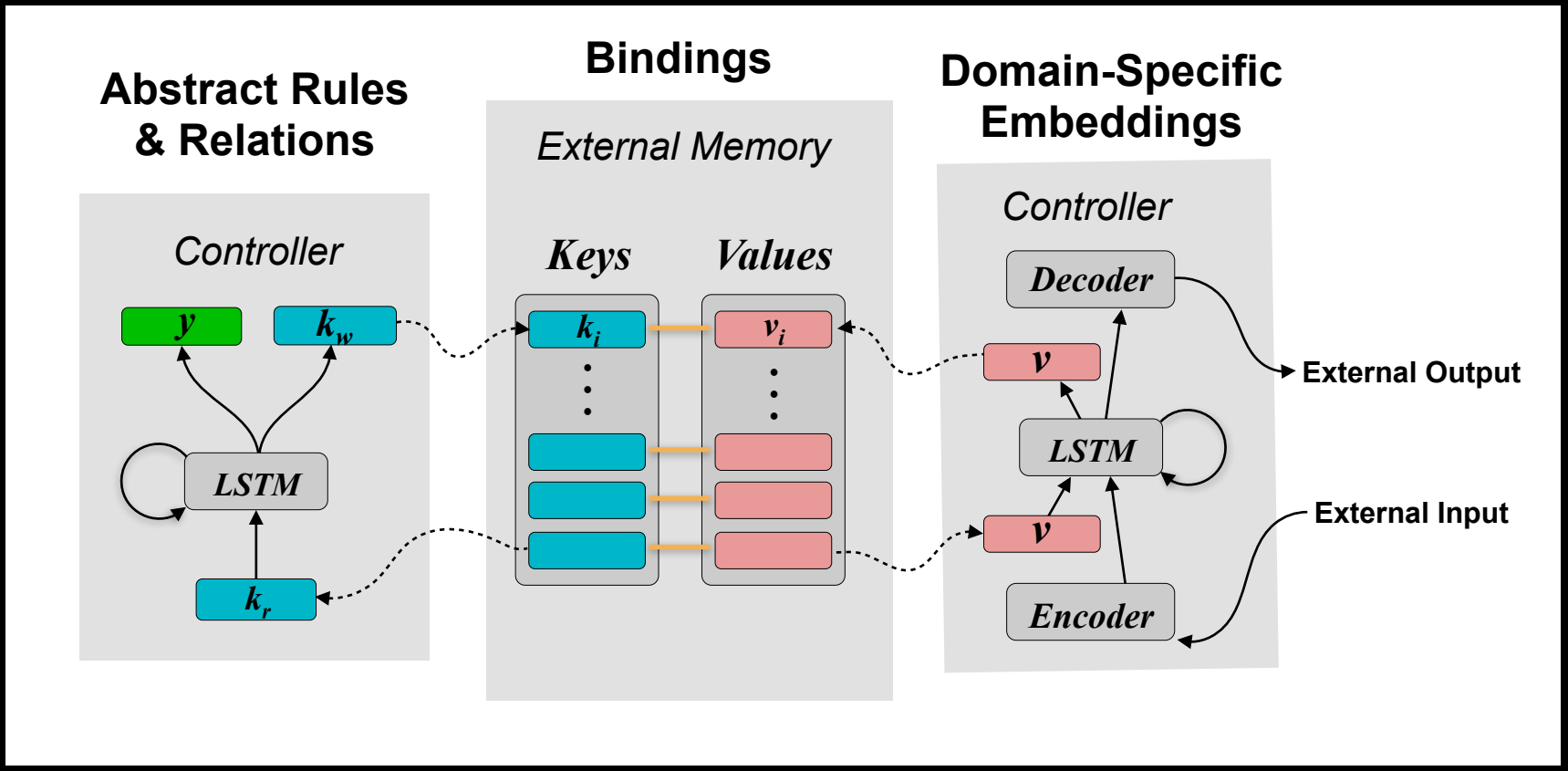
(Webb et al., ICLR 2021)



Neural Network with External Memory

Emergent Symbols Through Binding Network (ESBN)

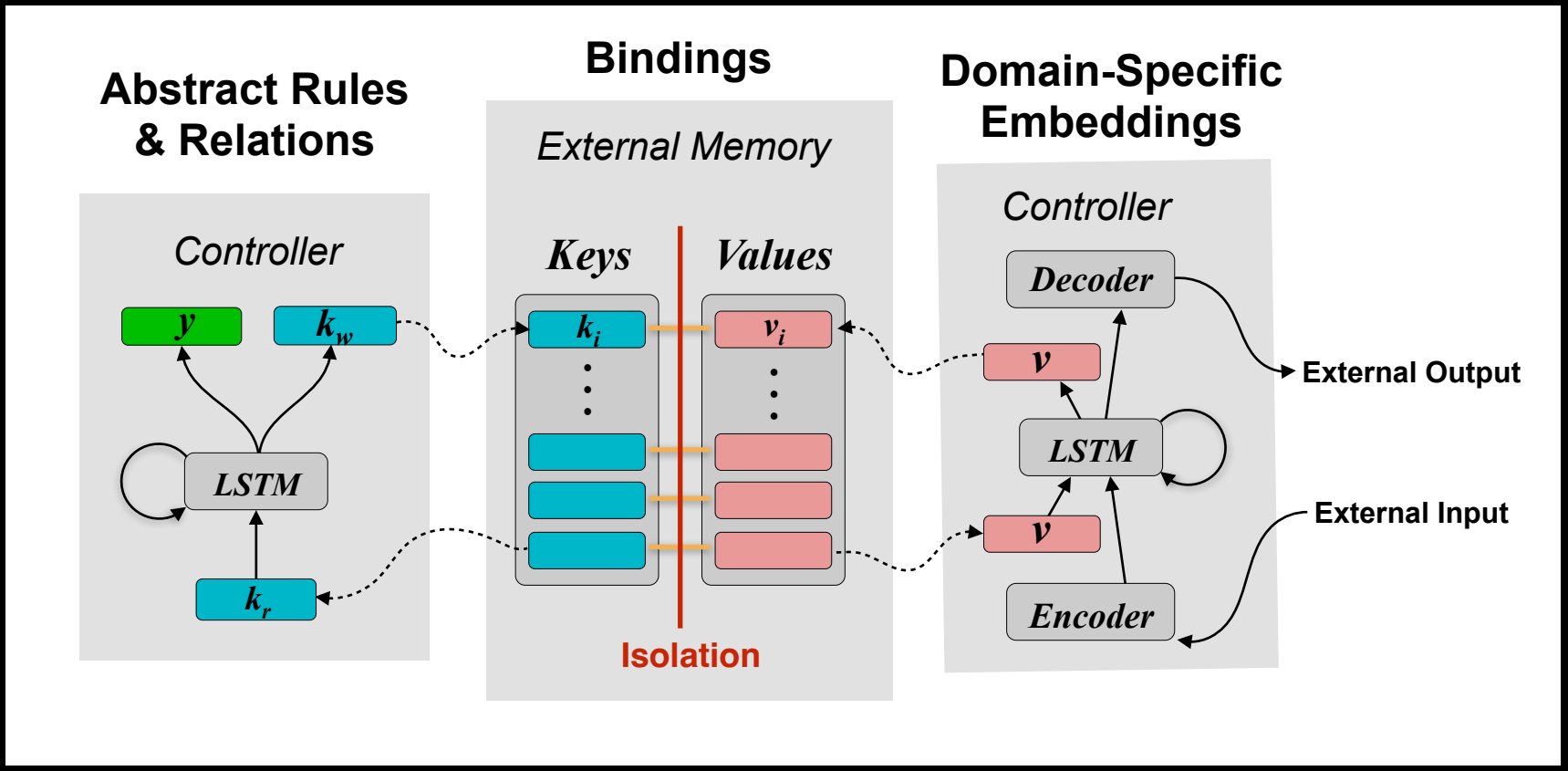
(Webb et al., ICLR 2021)



Neural Network with External Memory

Emergent Symbols Through Binding Network (ESBN)

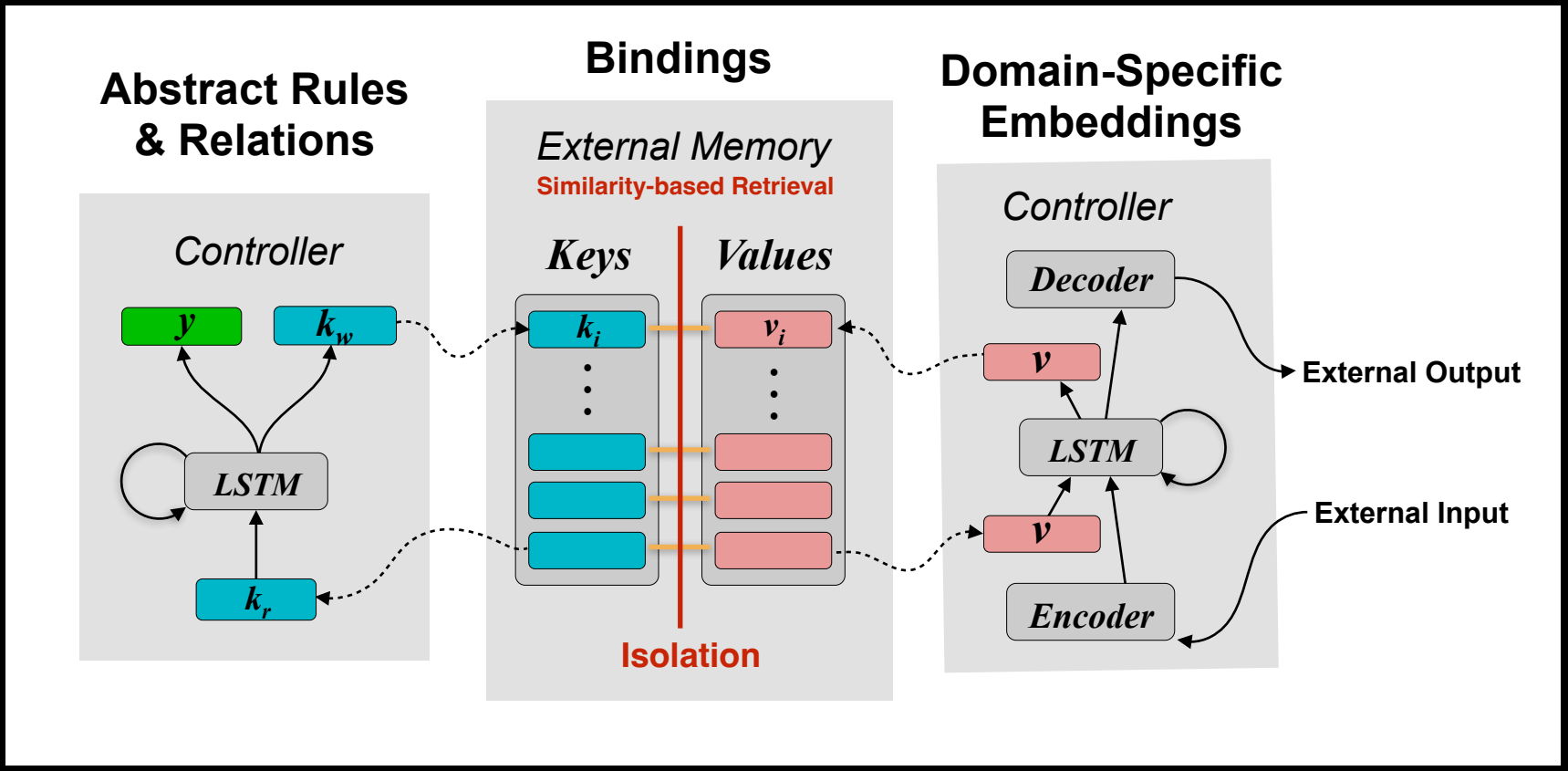
(Webb et al., ICLR 2021)



Relational Bottleneck

Emergent Symbols Through Binding Network (ESBN)

(Webb et al., ICLR 2021)



Isolation + similarity-based retrieval ⇒

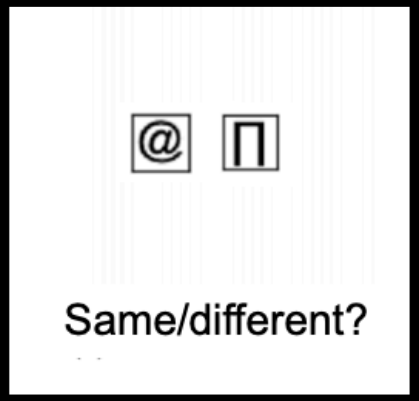
"Relational bottleneck"

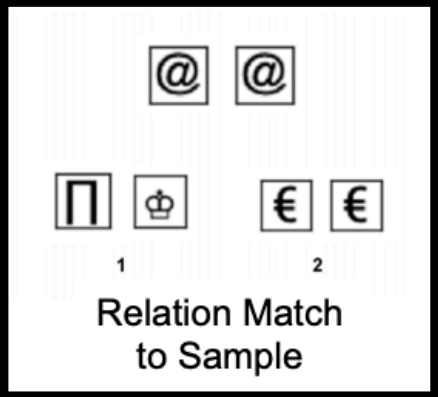
ESBN: Training

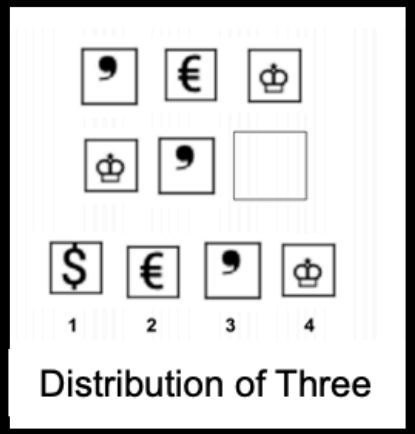
(Webb et al., 2021)

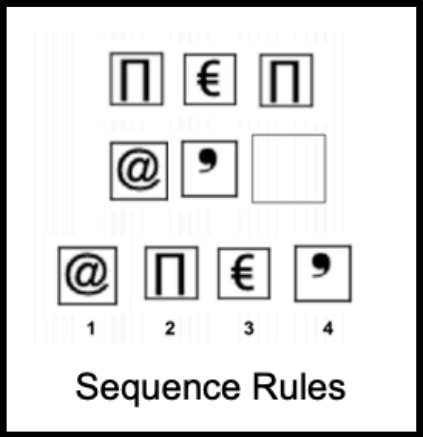
Simple Relational Tasks

from Ravens Progressive Matrices









ESBN: Training

(Webb et al., 2021)

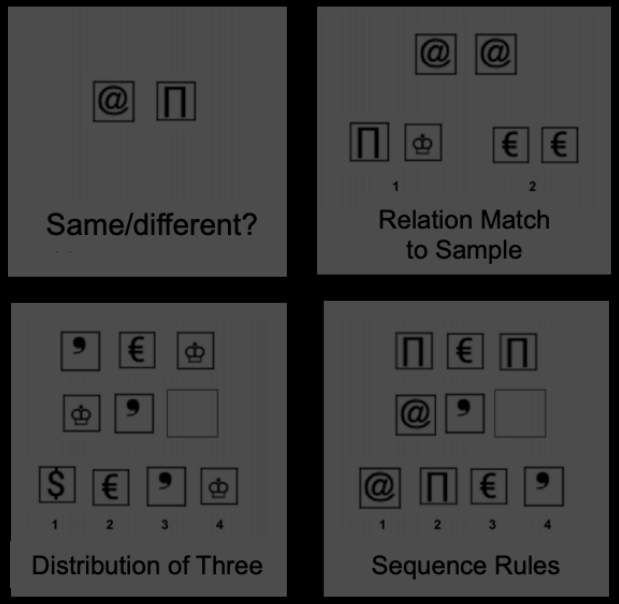
Extrapolation Performance

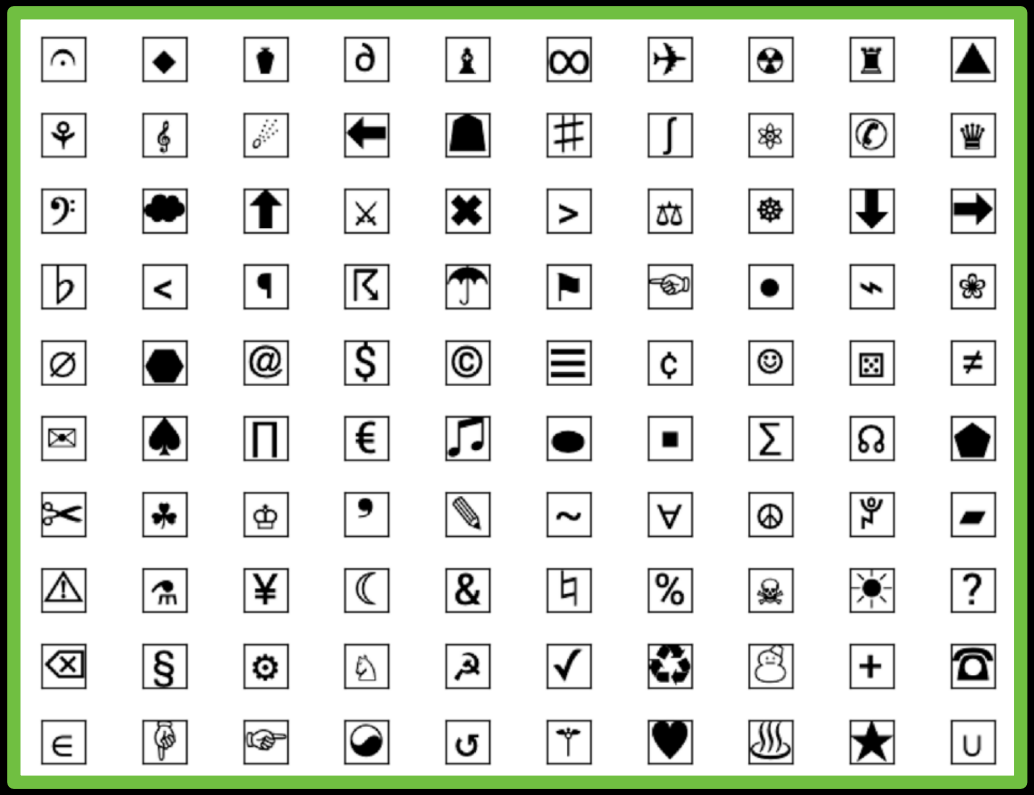
Trained on: *Tested on:*

100%

Simple Relational Tasks

from Ravens Progressive Matrices





ESBN: Training

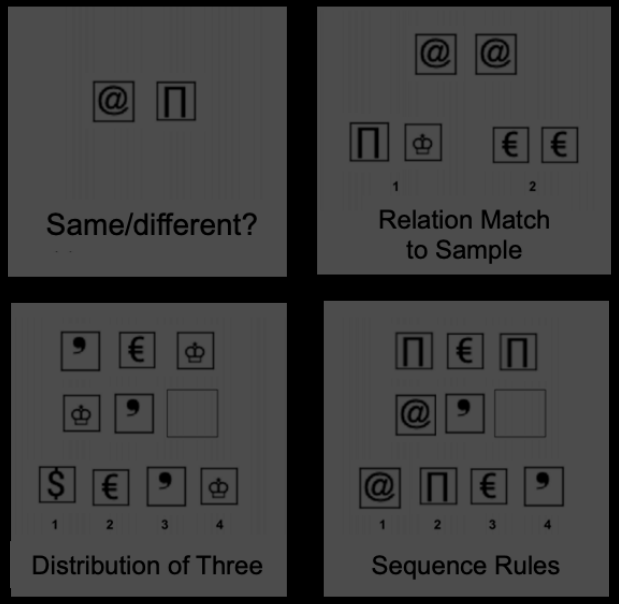
(Webb et al., 2021)

Extrapolation Performance

Trained on:

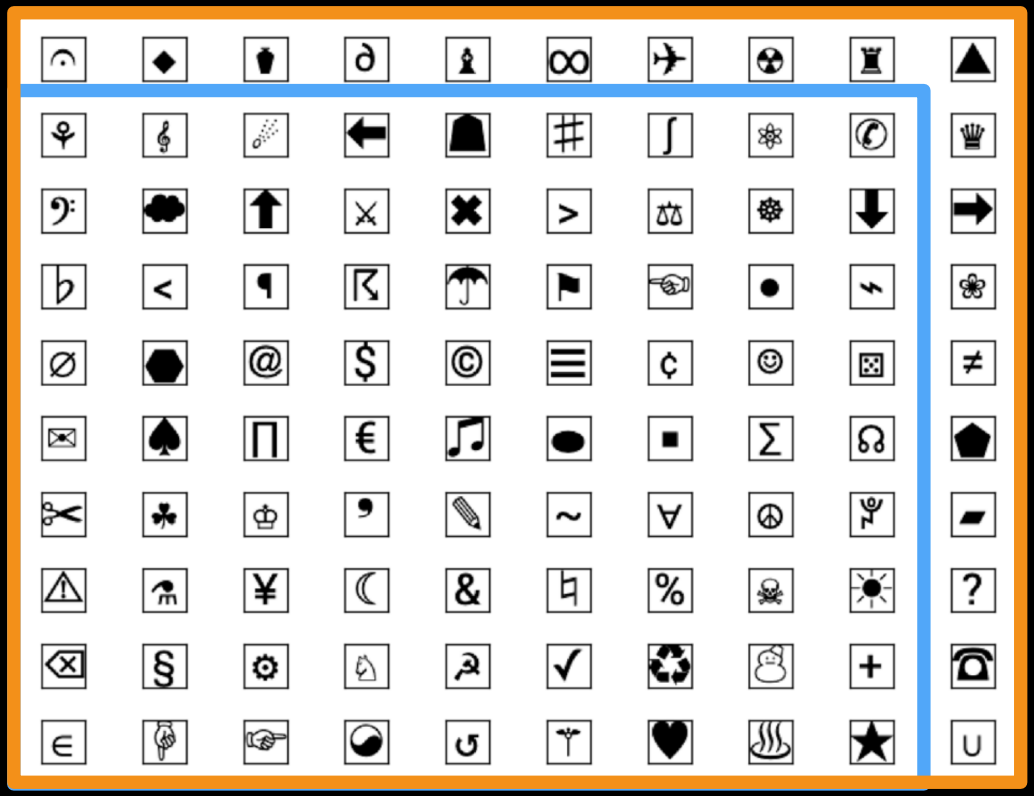
Simple Relational Tasks

from Ravens Progressive Matrices



85%

Tested on:



ESBN: Training

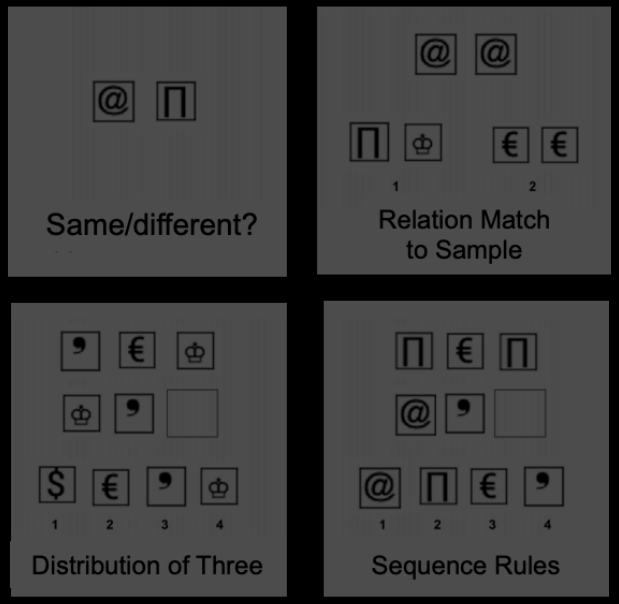
(Webb et al., 2021)

Extrapolation Performance

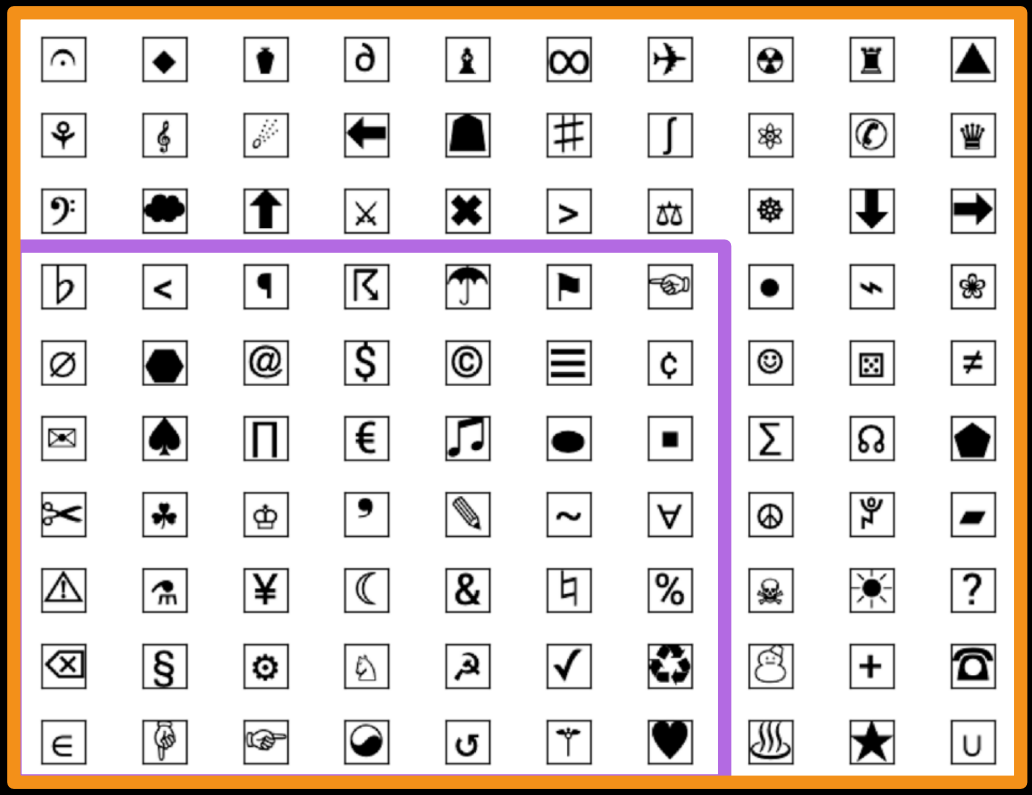
Trained on: *Tested on:*

Simple Relational Tasks

from Ravens Progressive Matrices



50%



ESBN: Training

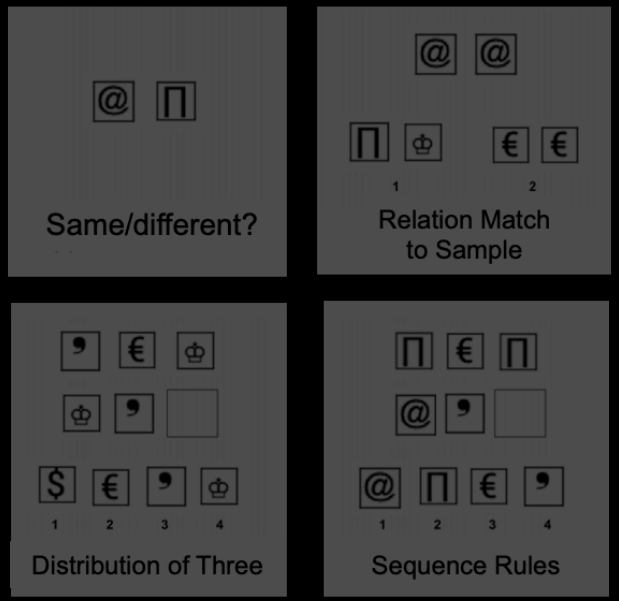
(Webb et al., 2021)

Extrapolation Performance

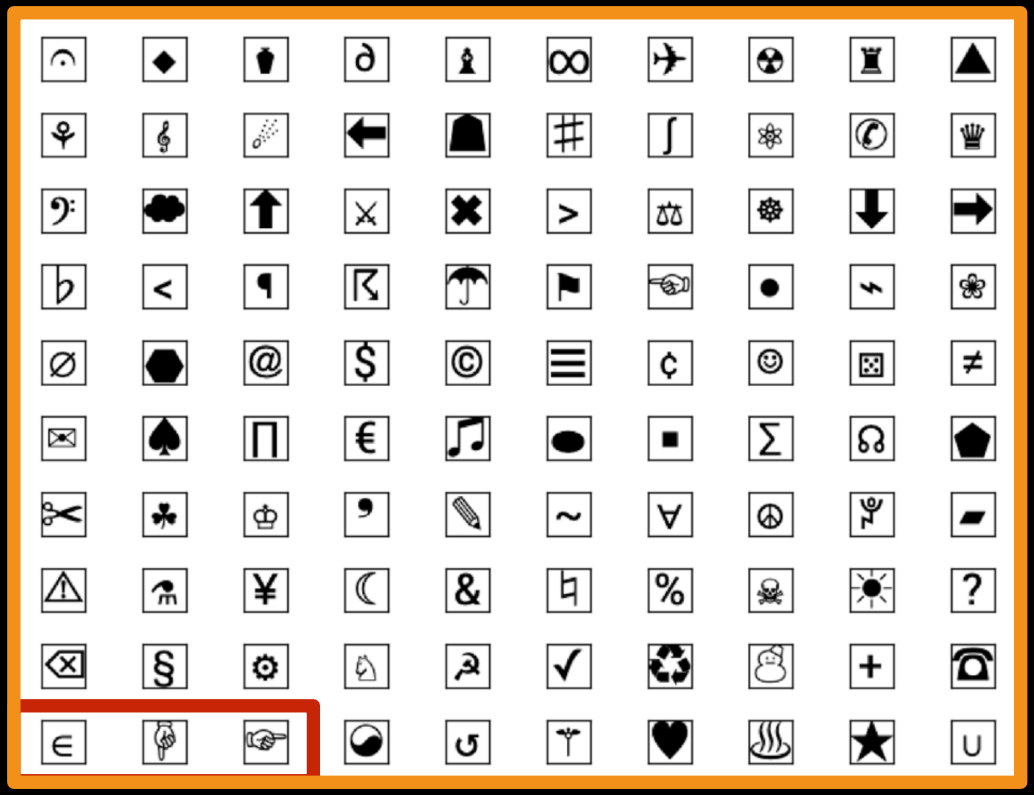
Trained on:

Simple Relational Tasks

from Ravens Progressive Matrices



Tested on:

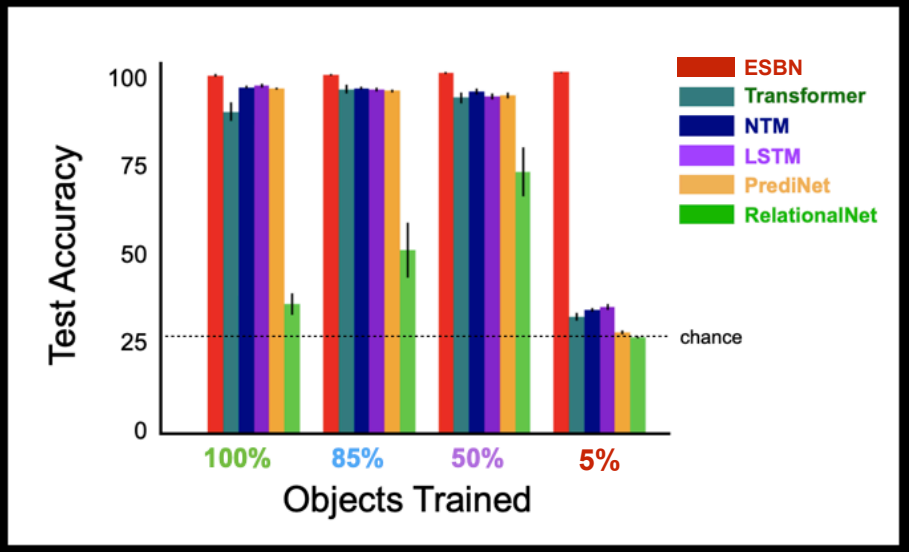


5%

ESBN: Results

(Webb et al., ICLR 2021)

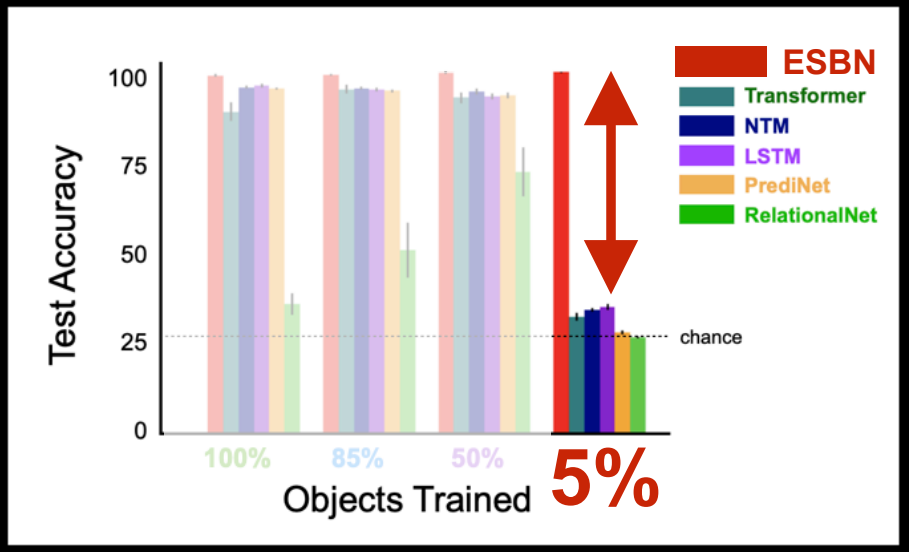
Out of Range Extrapolation



ESBN: Results

(Webb et al., ICLR 2021)

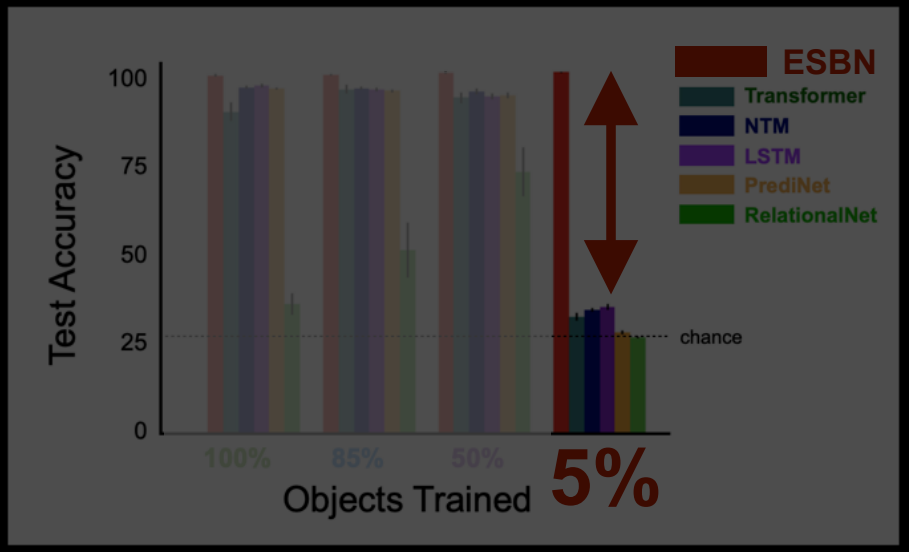
Out of Range Extrapolation



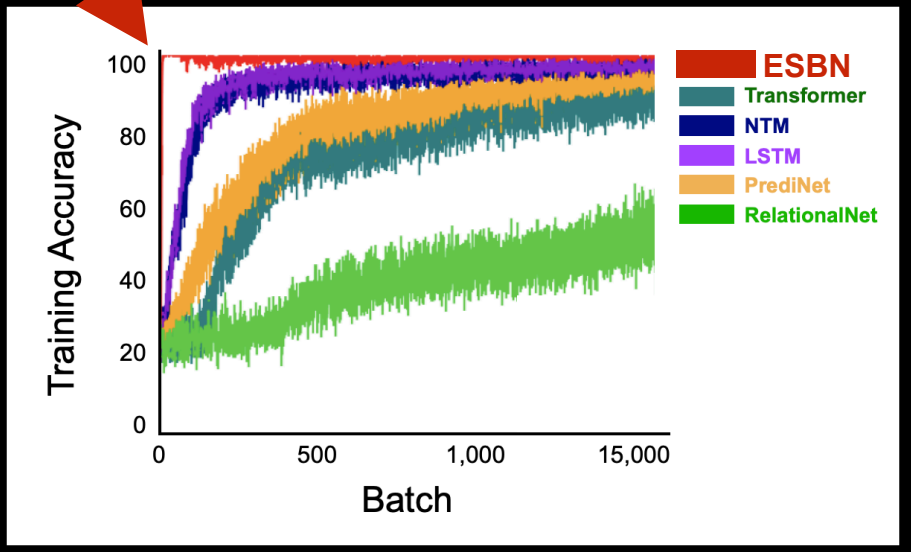
ESBN: Results

(Webb et al., ICLR 2021)

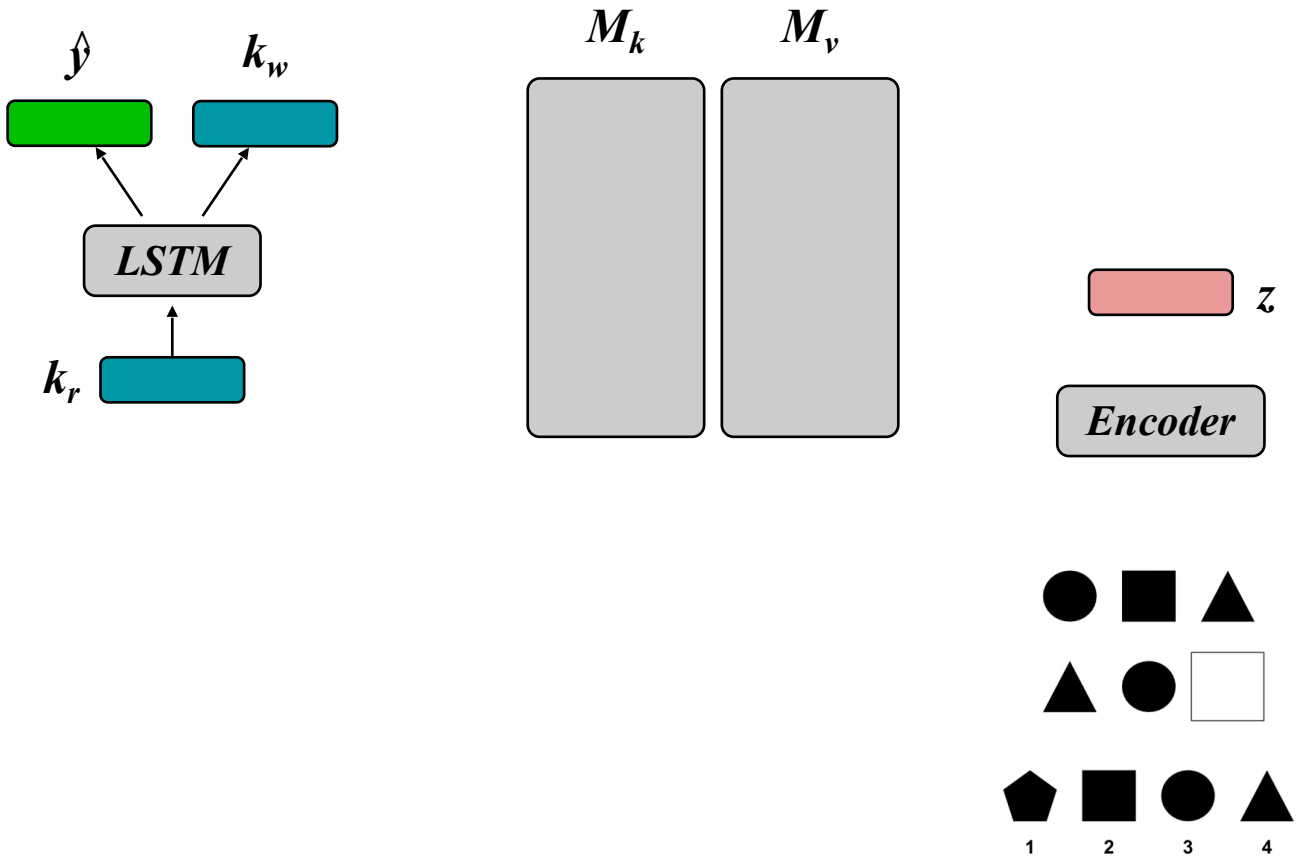
Out of Range Extrapolation



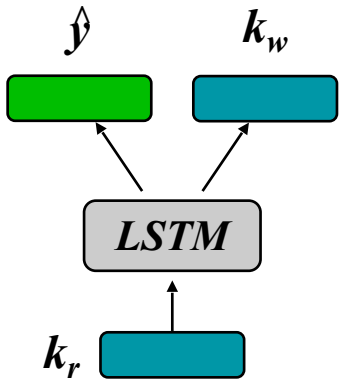
Sample Efficiency

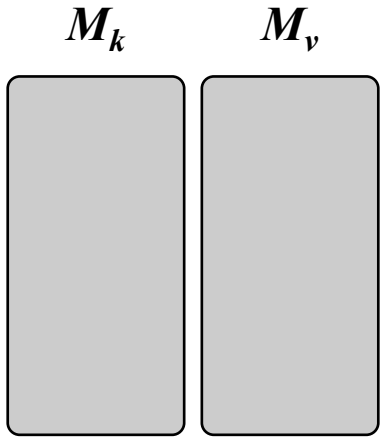


Example Trial: Distribution of Three

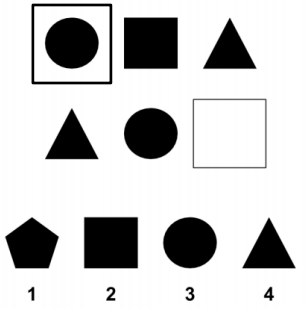


Example Trial: Distribution of Three

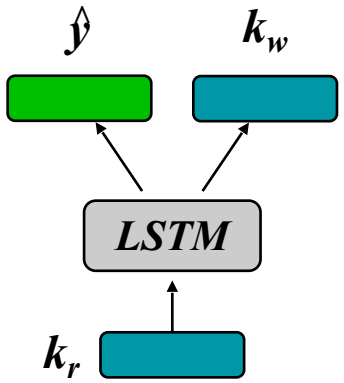


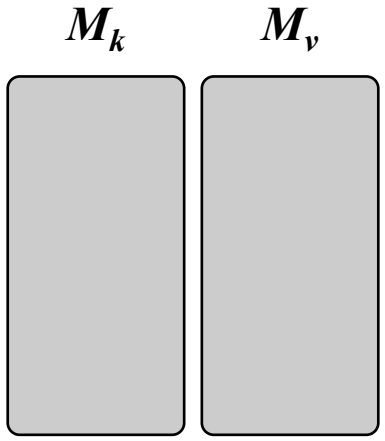


Encode

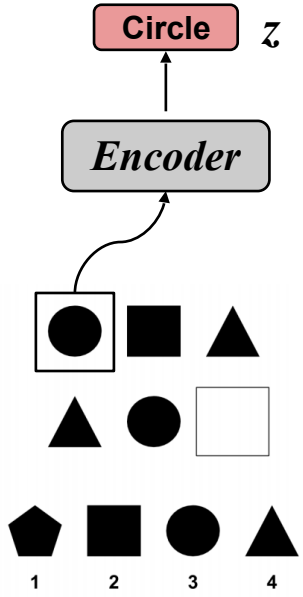


Example Trial: Distribution of Three

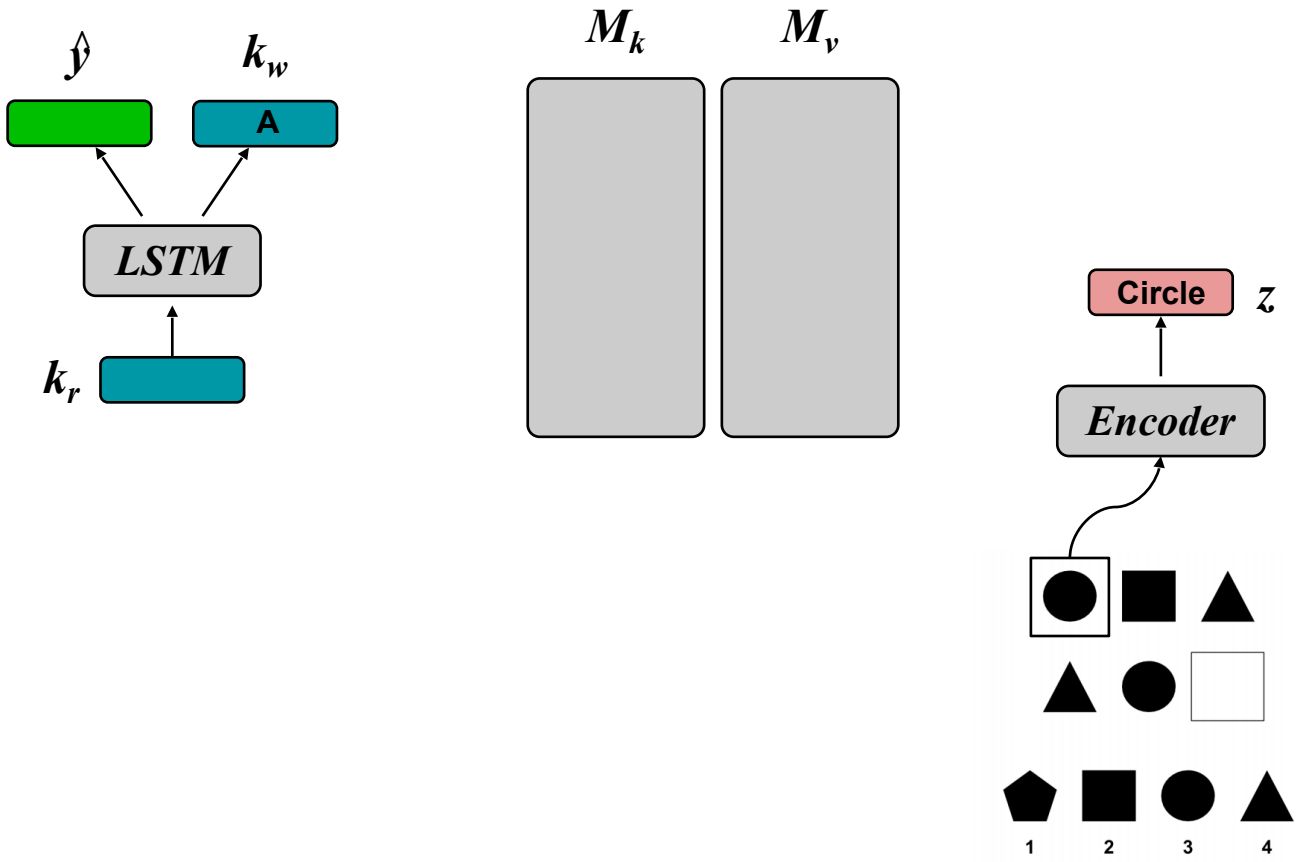




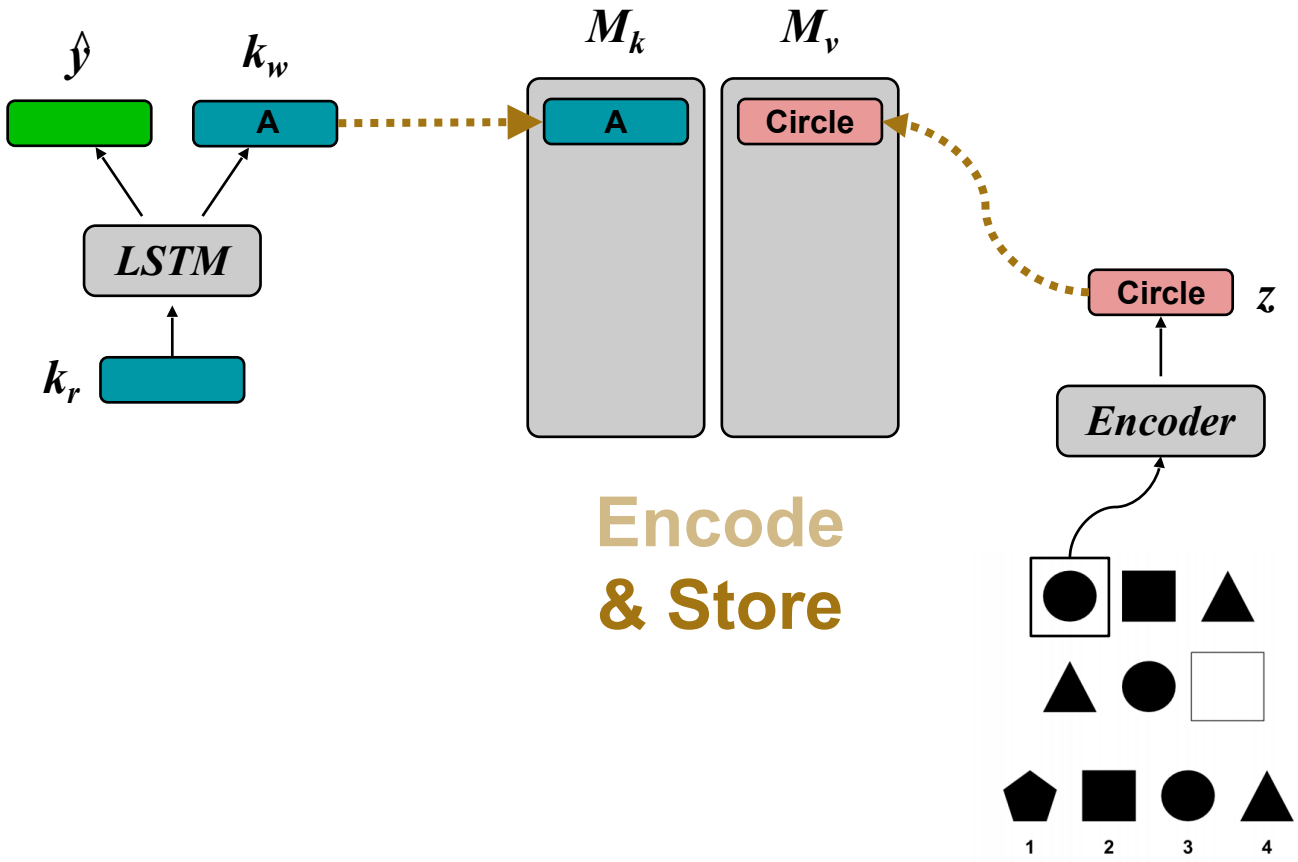
Encode



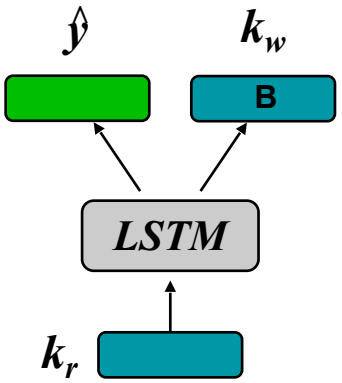
Example Trial: Distribution of Three

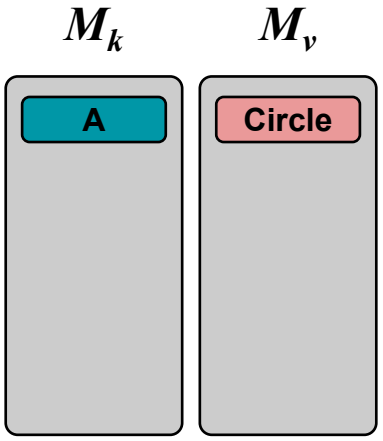


Example Trial: Distribution of Three



Example Trial: Distribution of Three

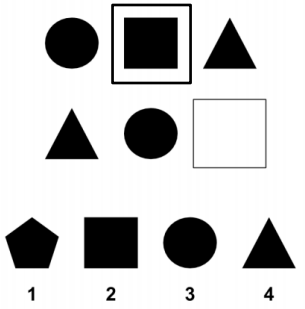




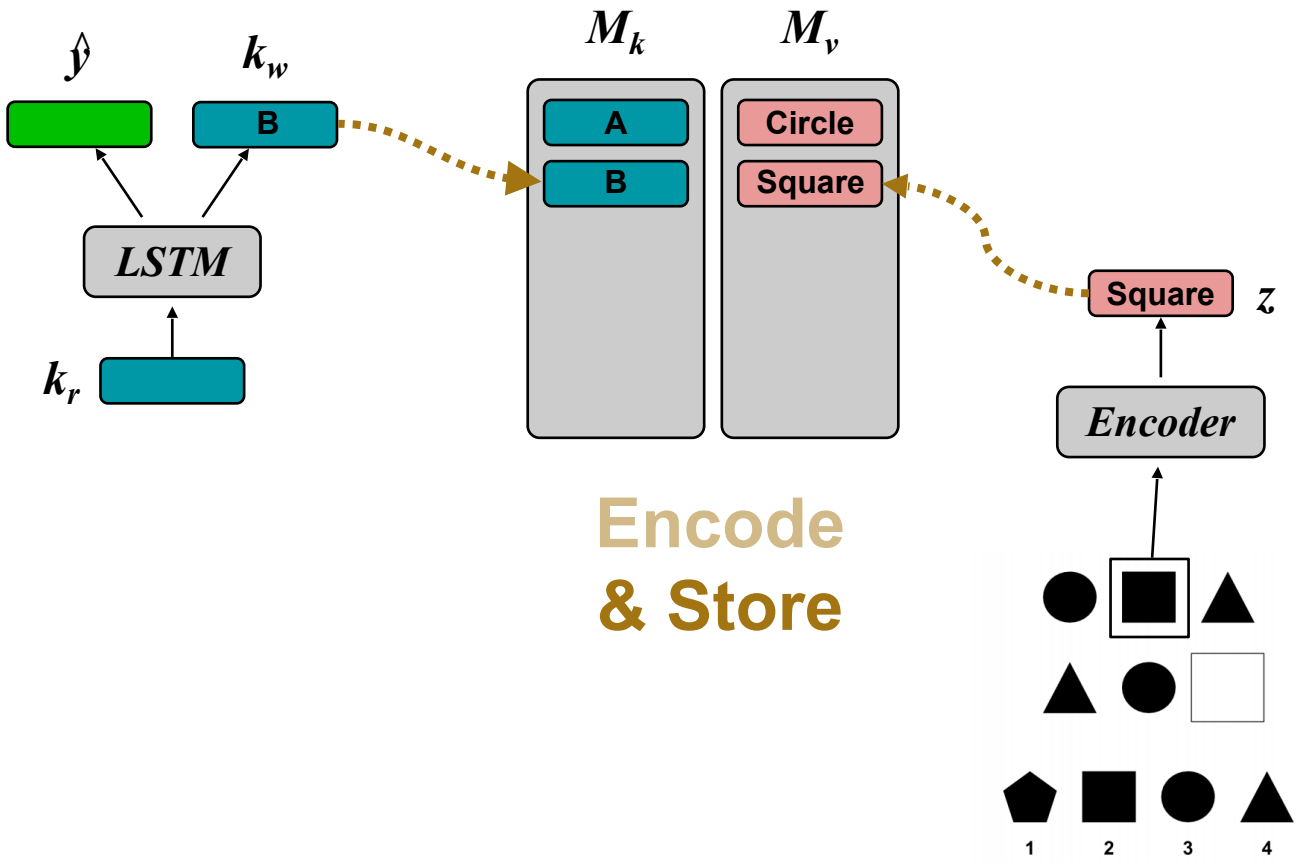
Encode
& Store

Square z

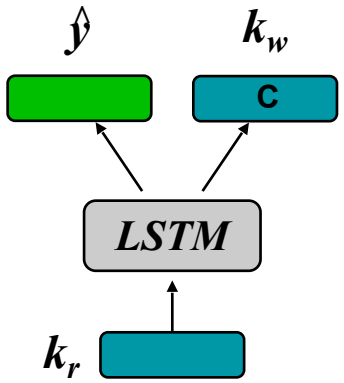
Encoder

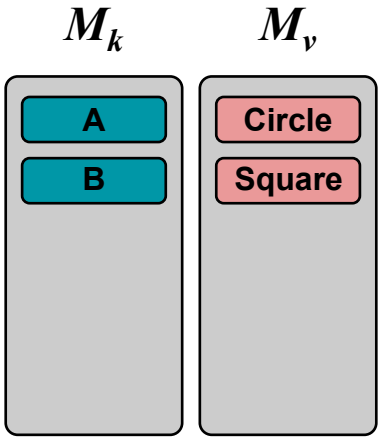


Example Trial: Distribution of Three

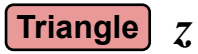


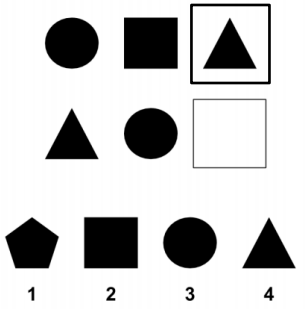
Example Trial: Distribution of Three



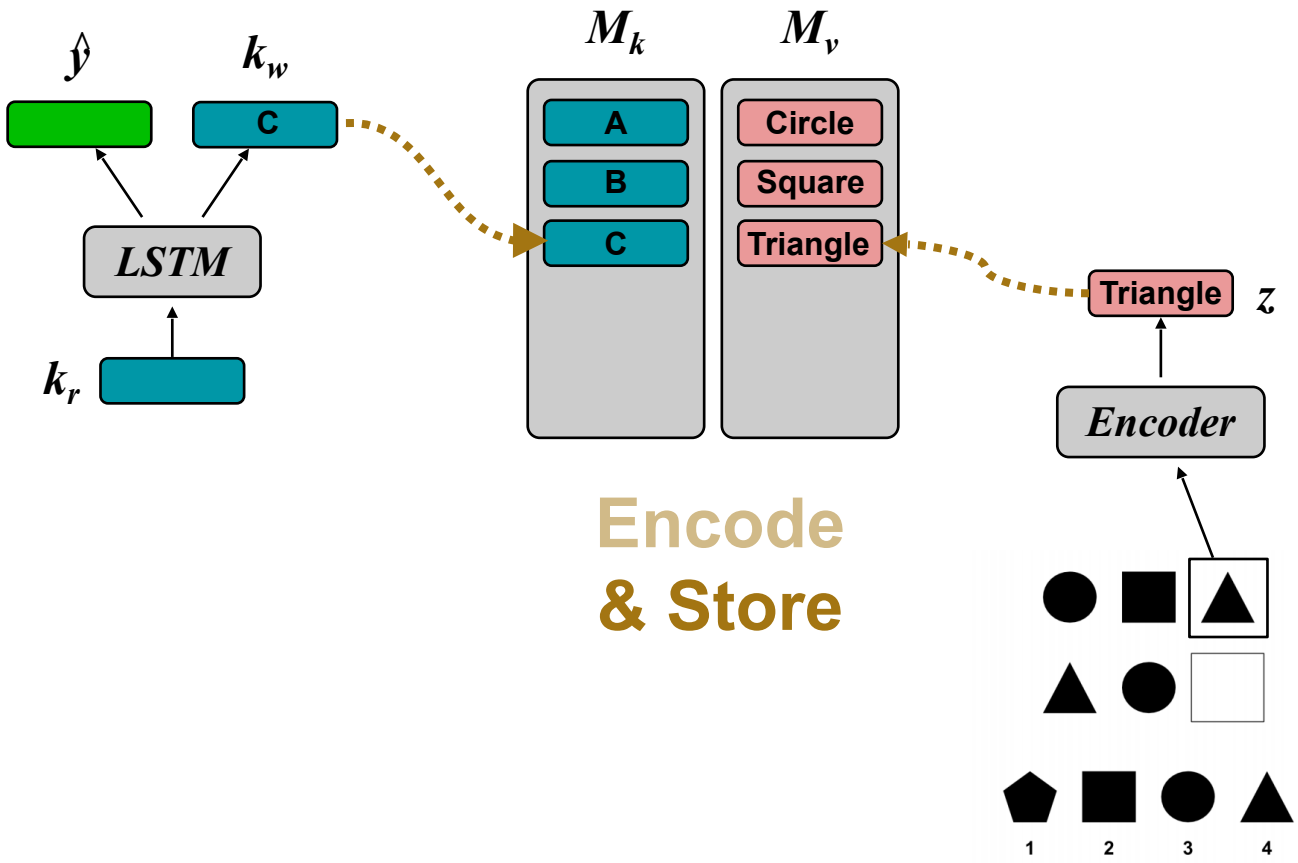


Encode
& Store

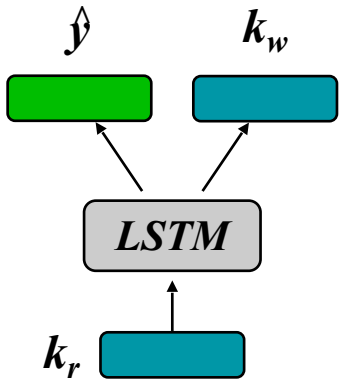


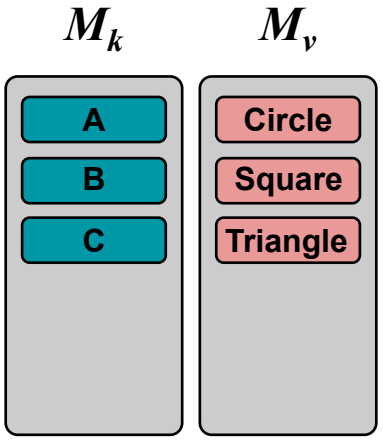


Example Trial: Distribution of Three

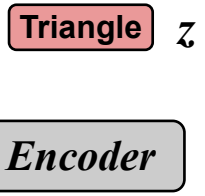


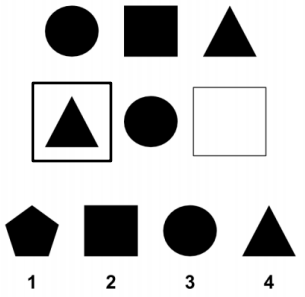
Example Trial: Distribution of Three



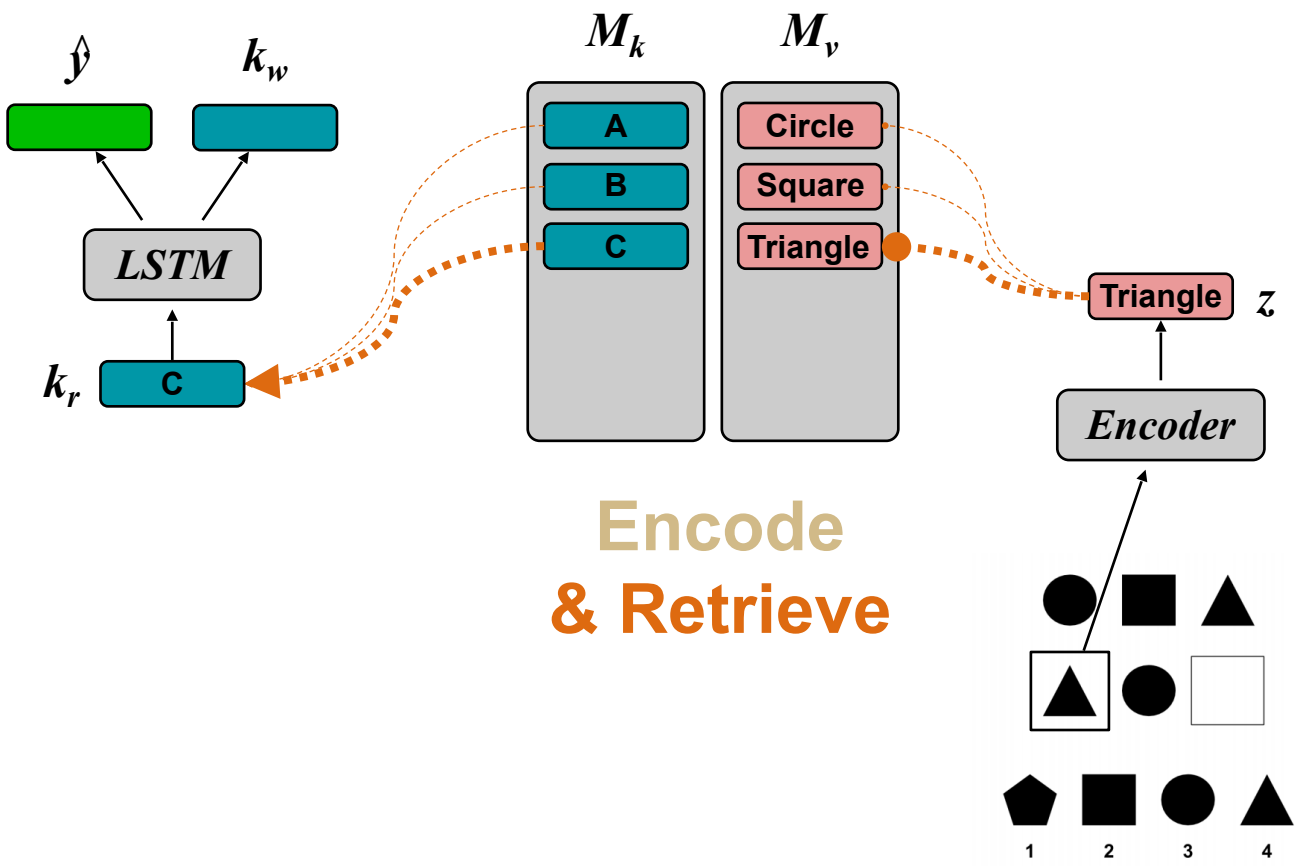


Encode
& Retrieve

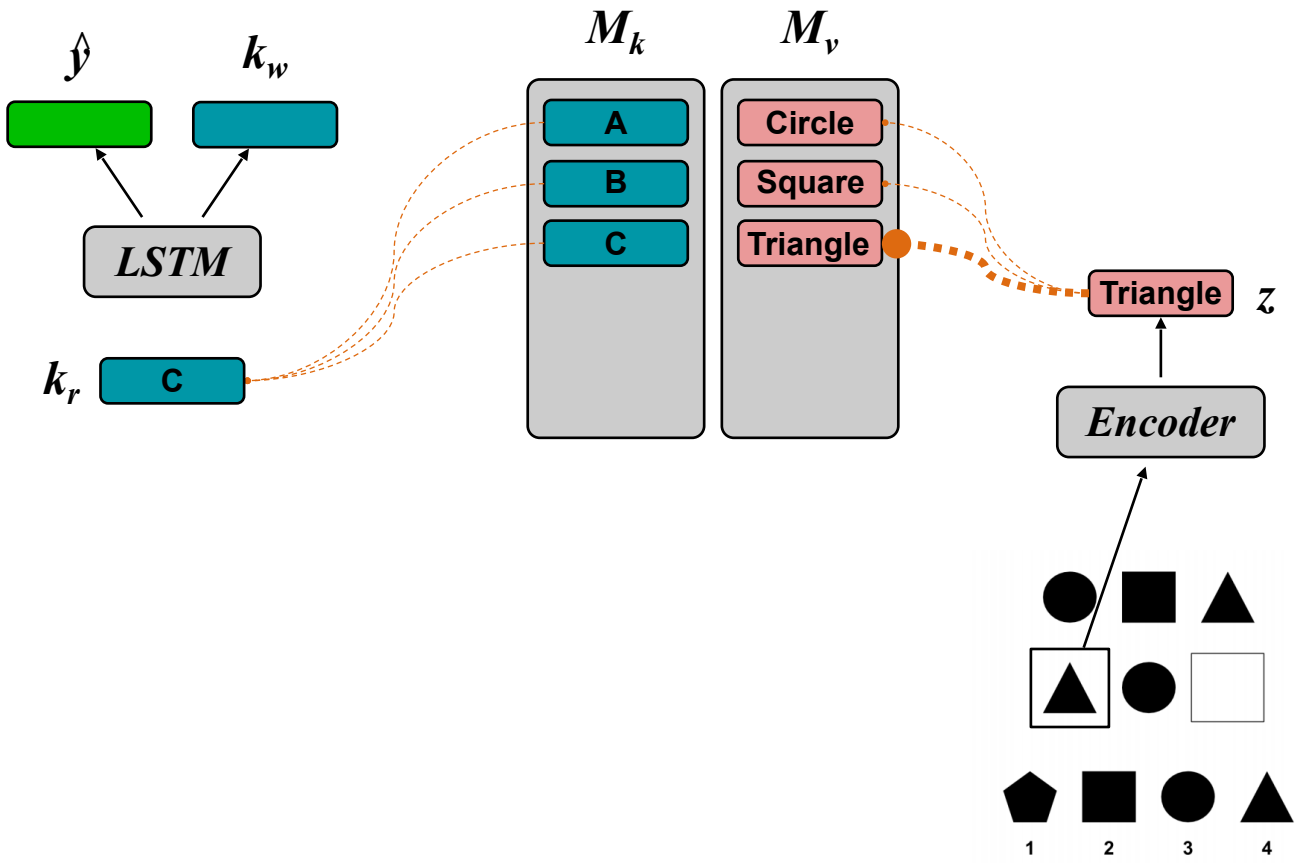




Example Trial: Distribution of Three

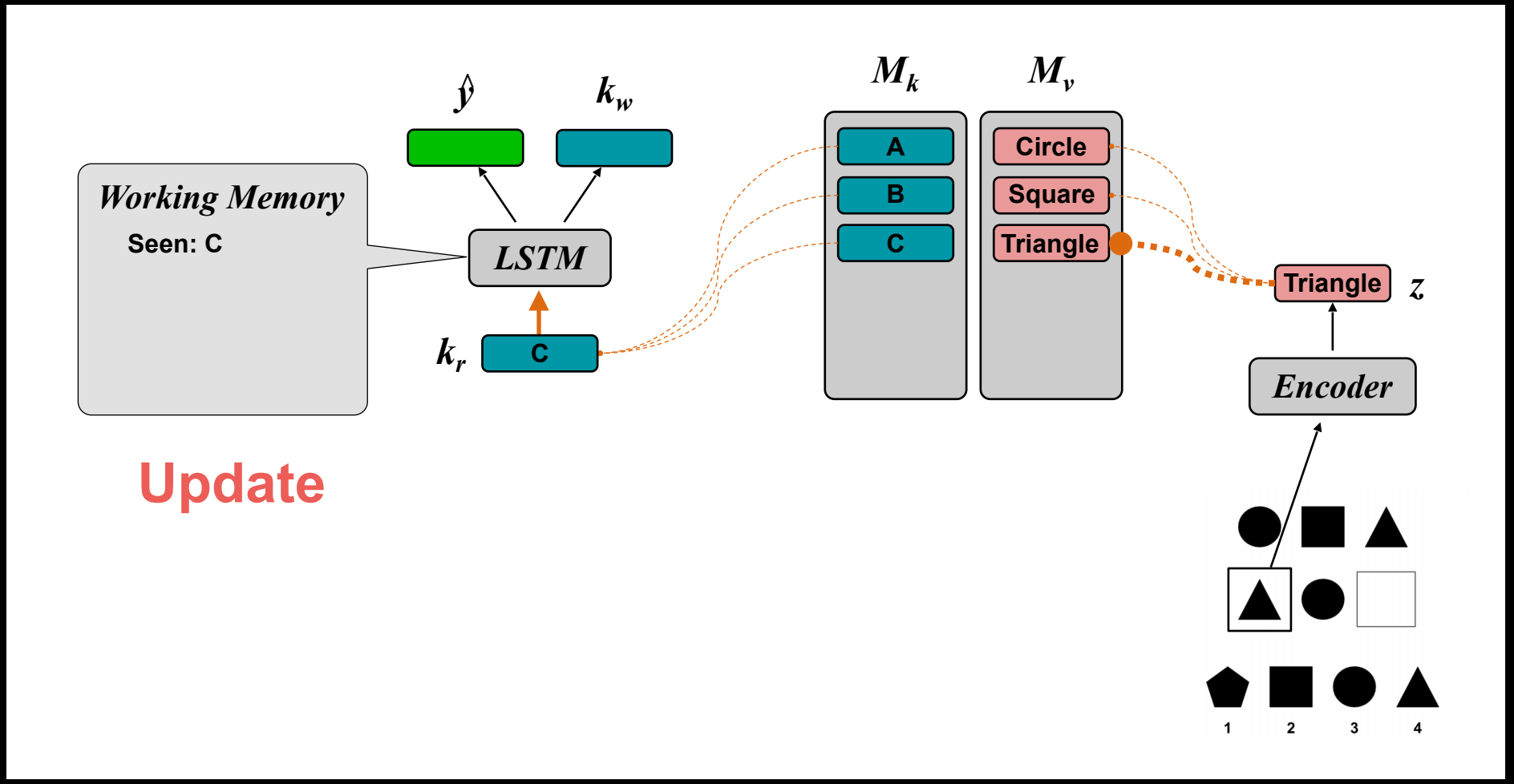


Example Trial: Distribution of Three

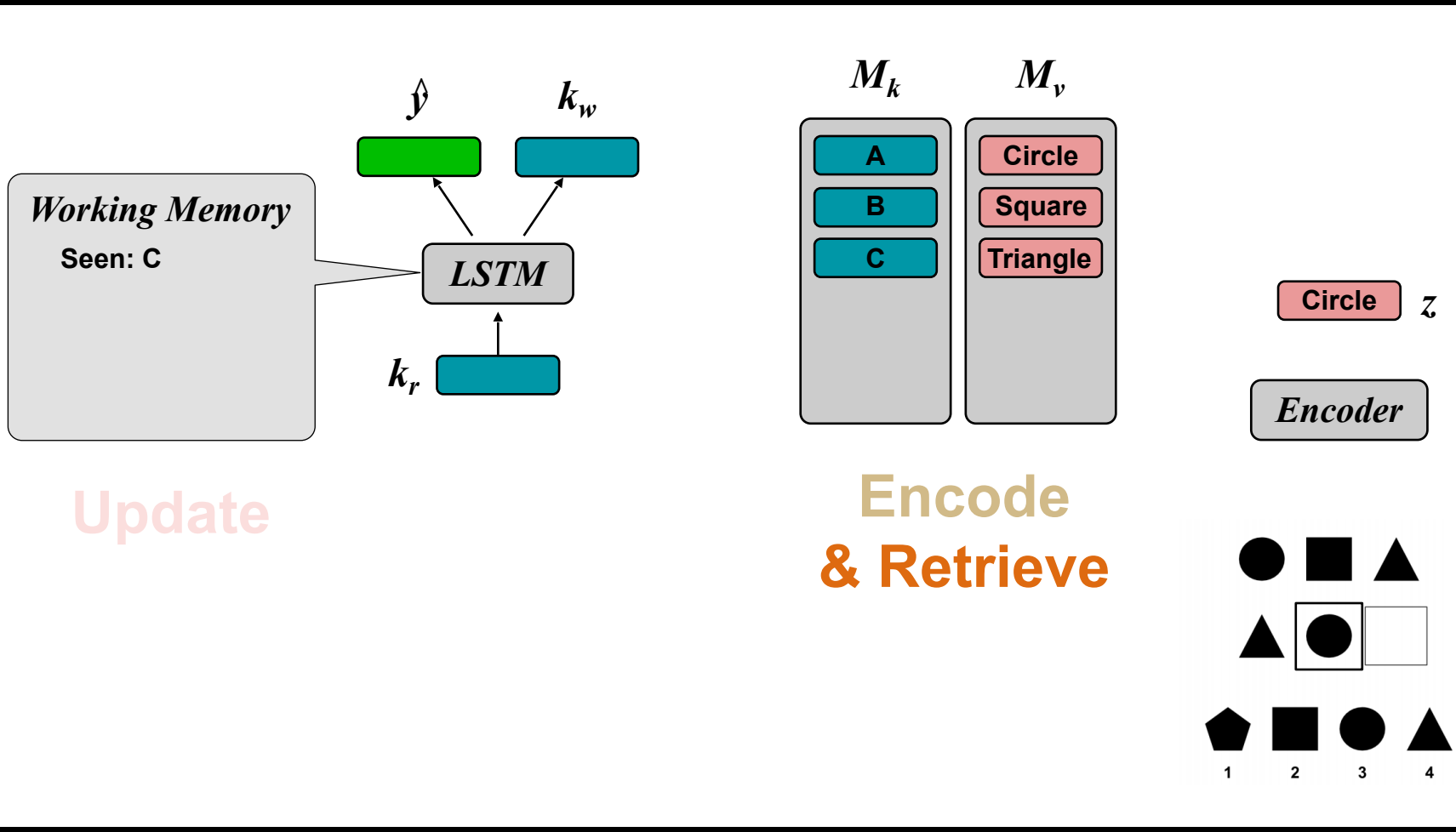


Update

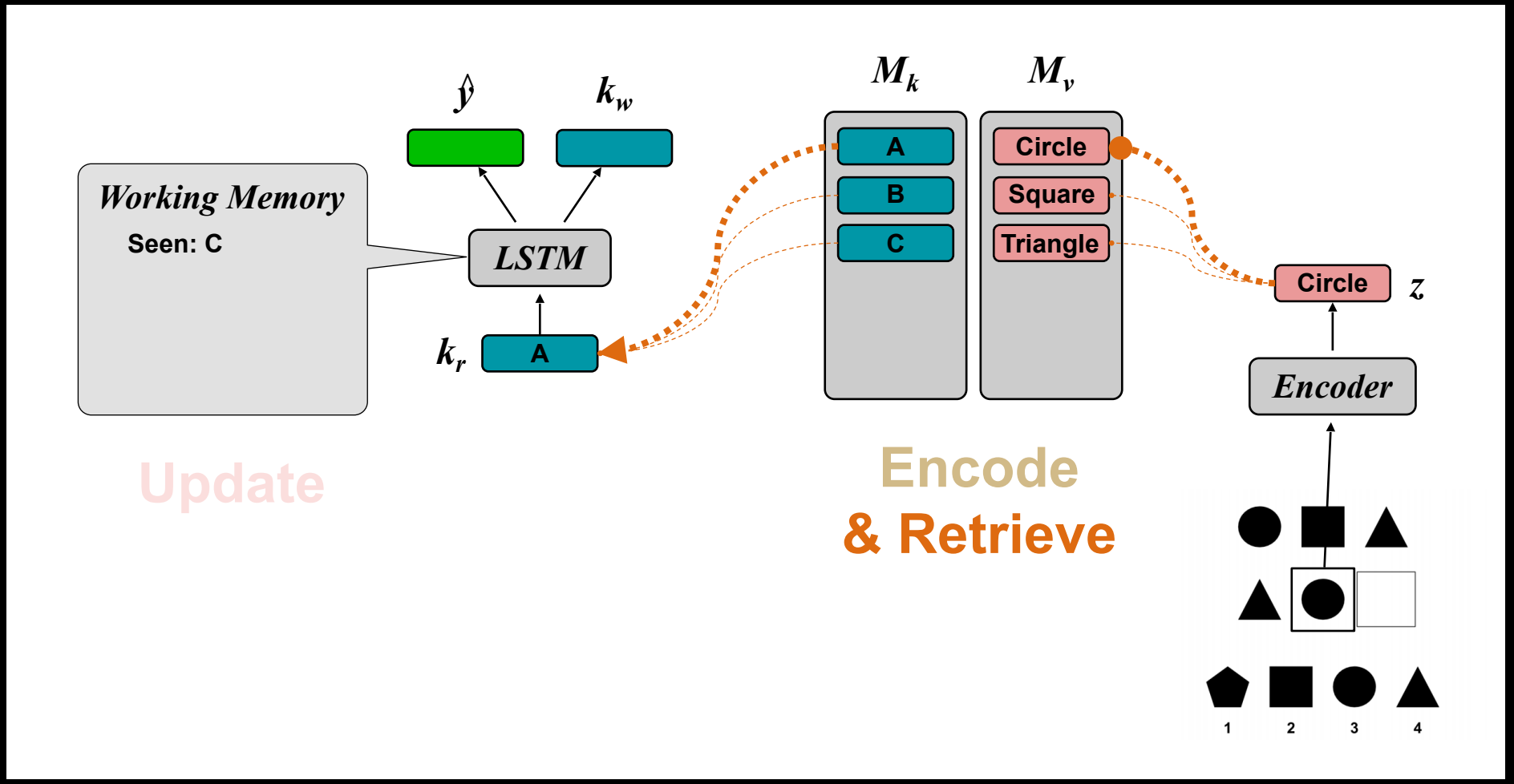
Example Trial: Distribution of Three



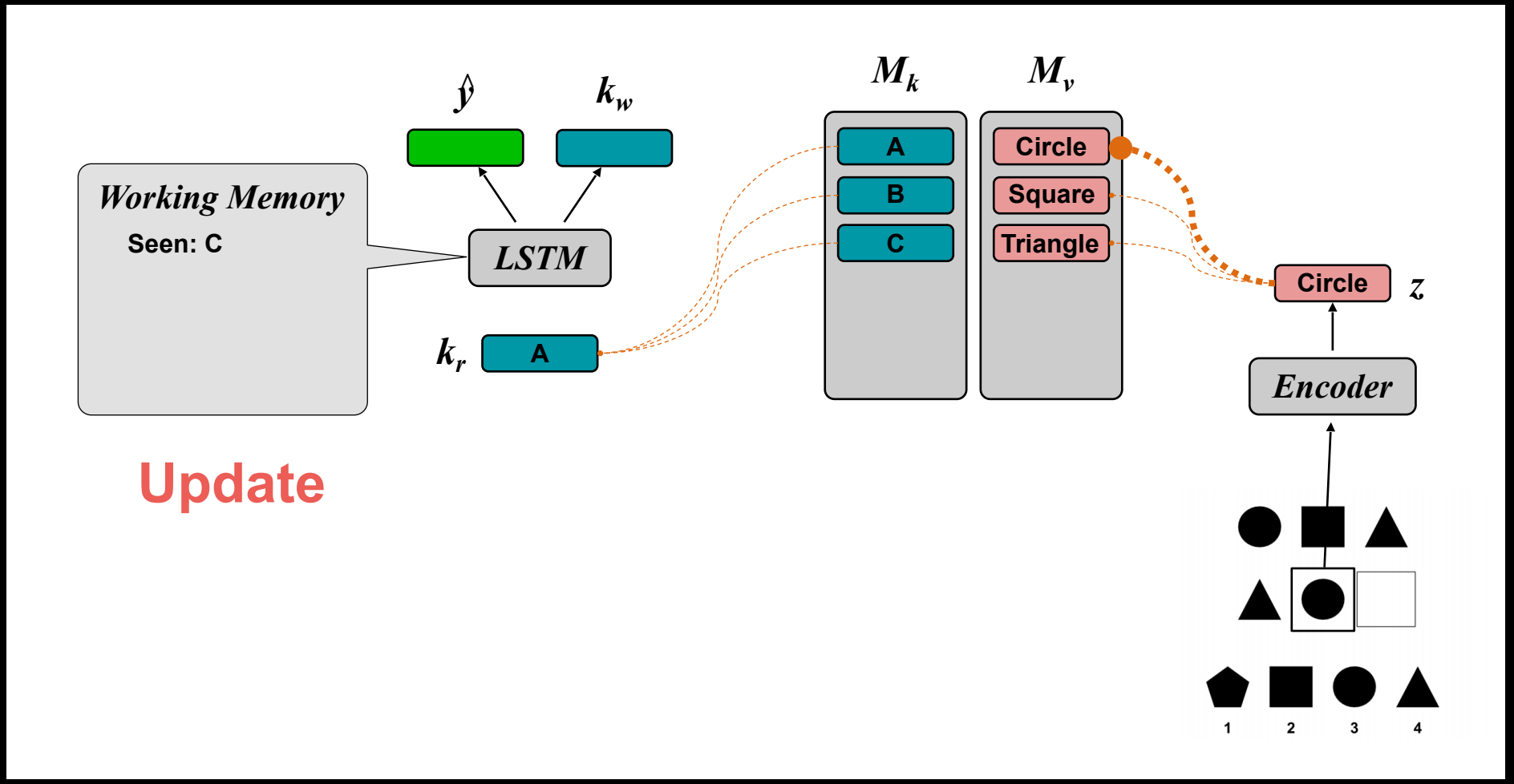
Example Trial: Distribution of Three



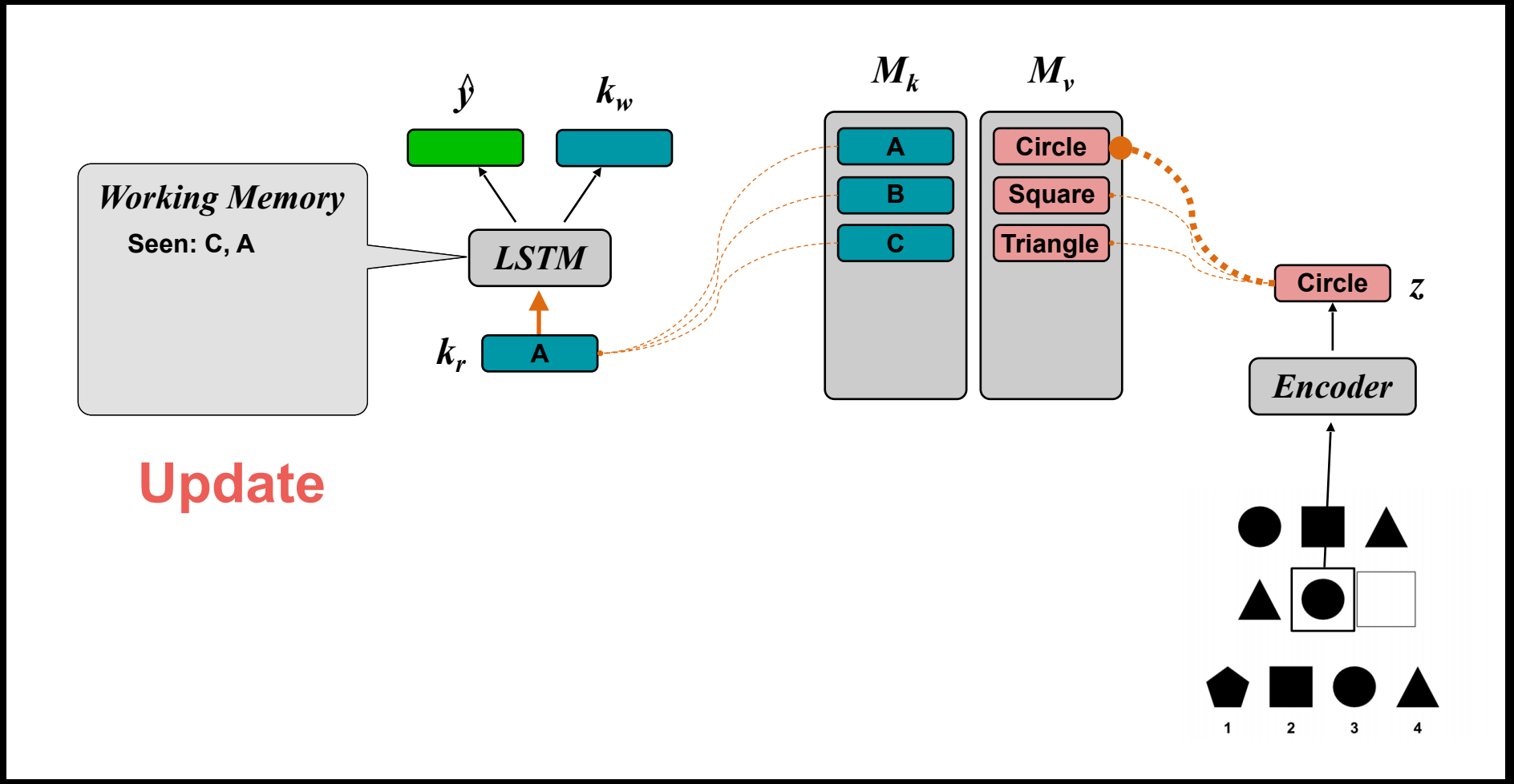
Example Trial: Distribution of Three



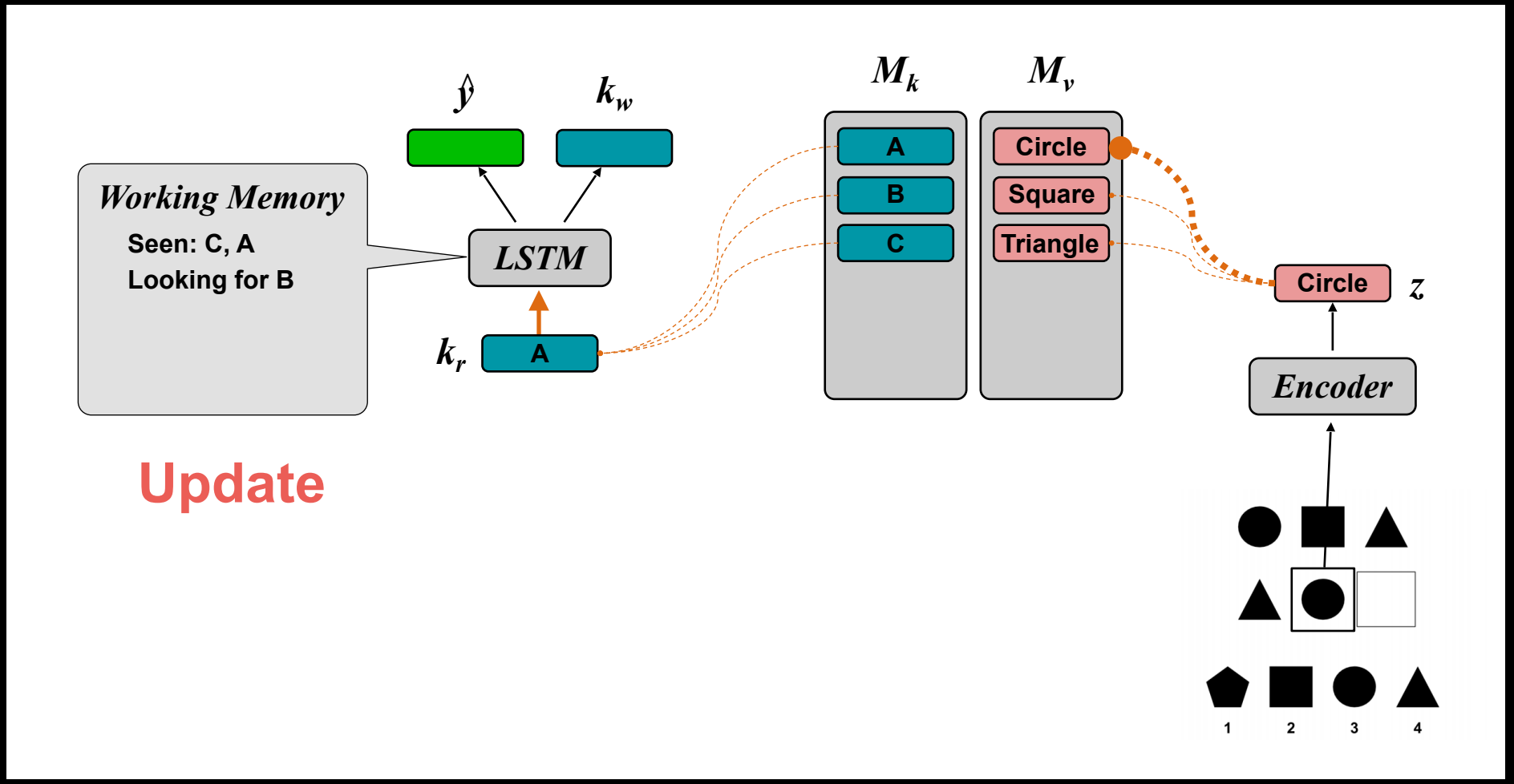
Example Trial: Distribution of Three



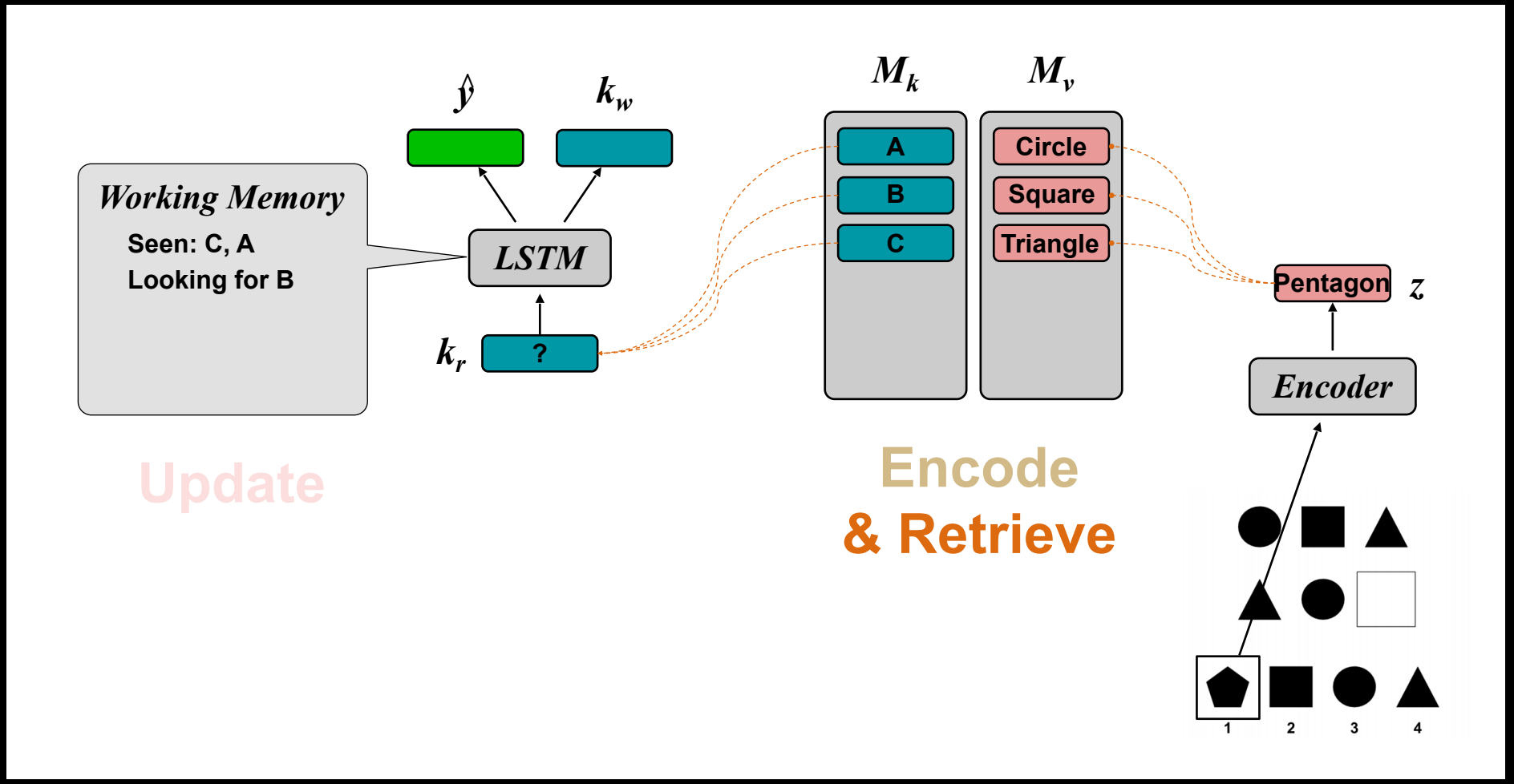
Example Trial: Distribution of Three



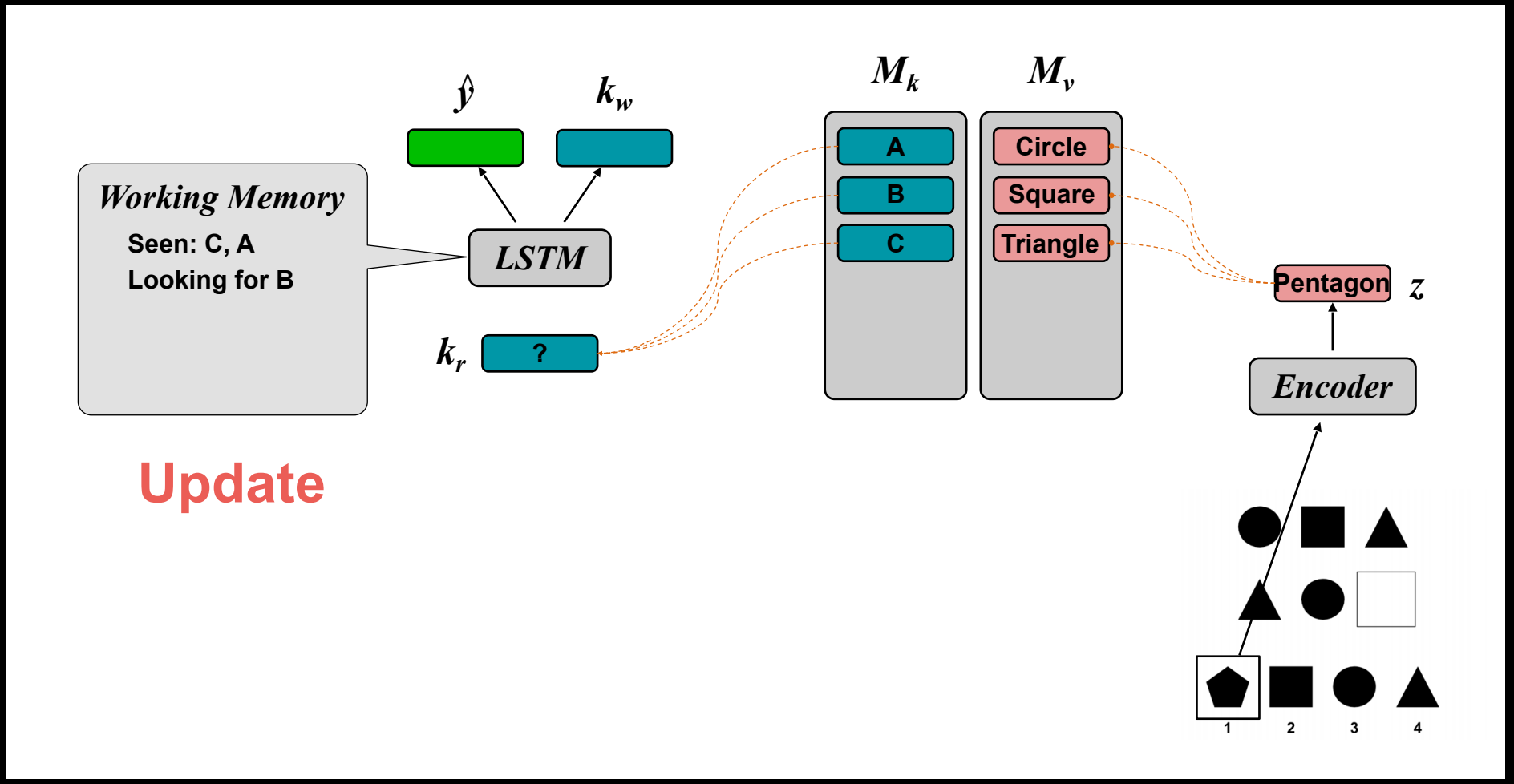
Example Trial: Distribution of Three



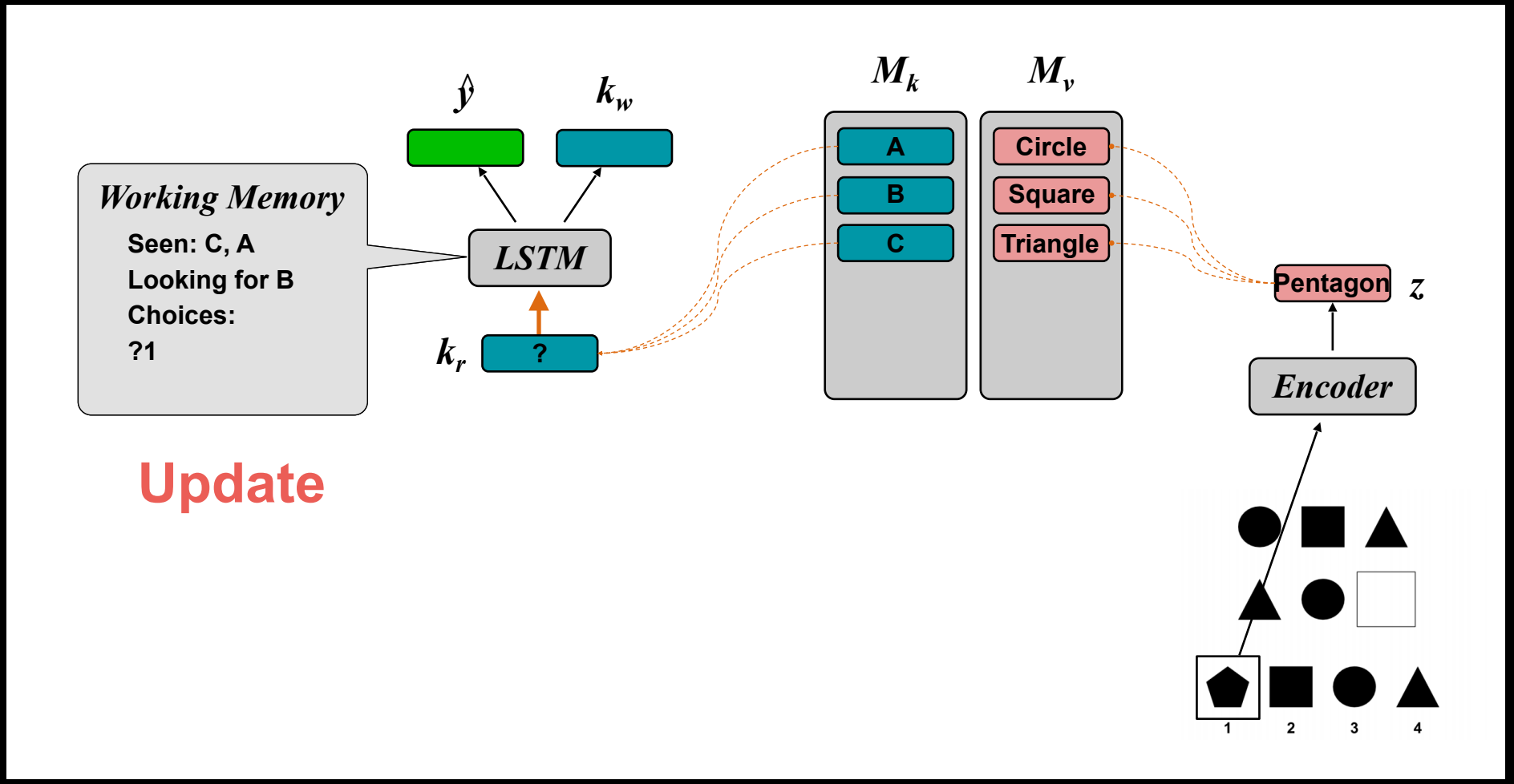
Example Trial: Distribution of Three



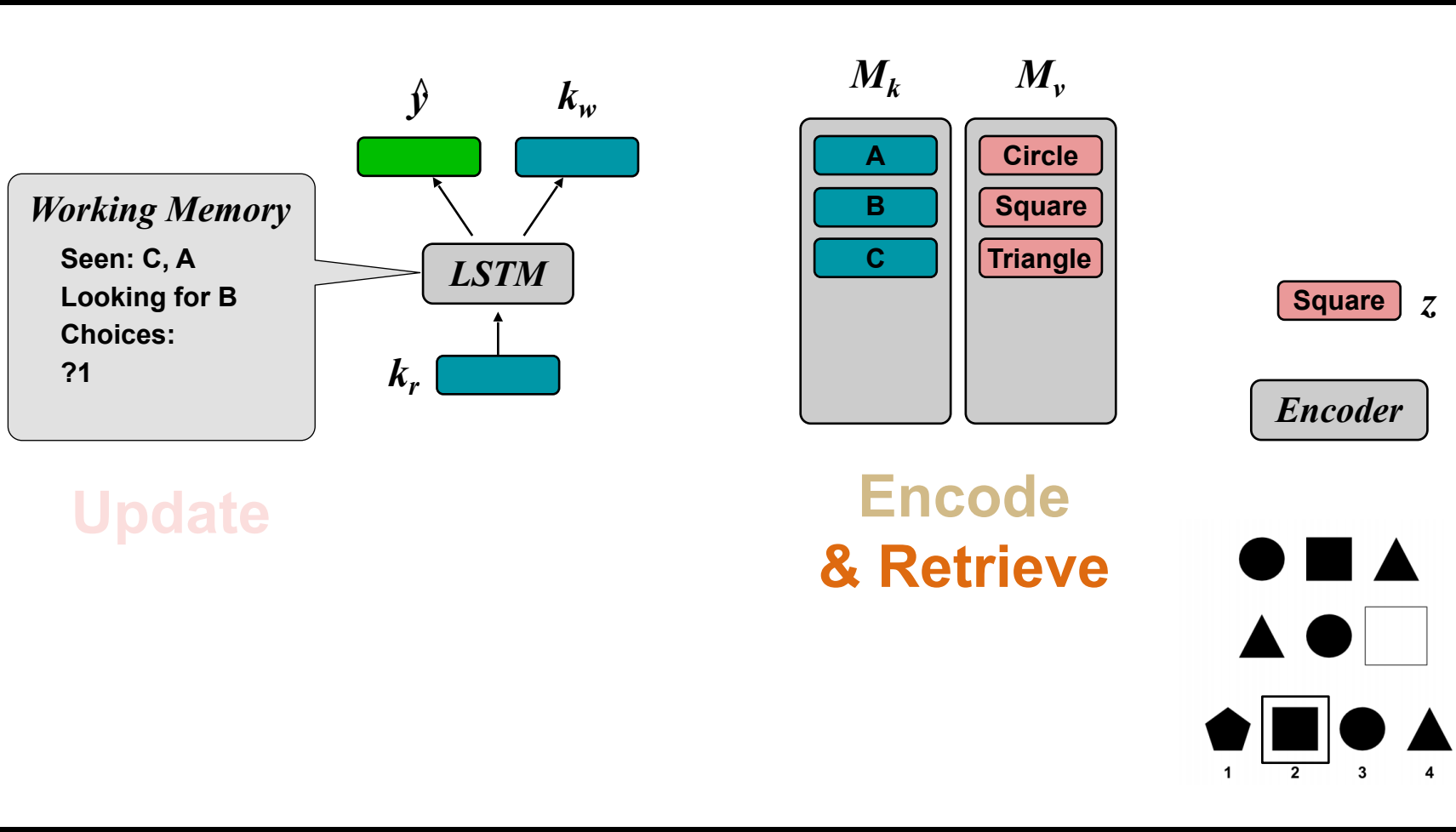
Example Trial: Distribution of Three



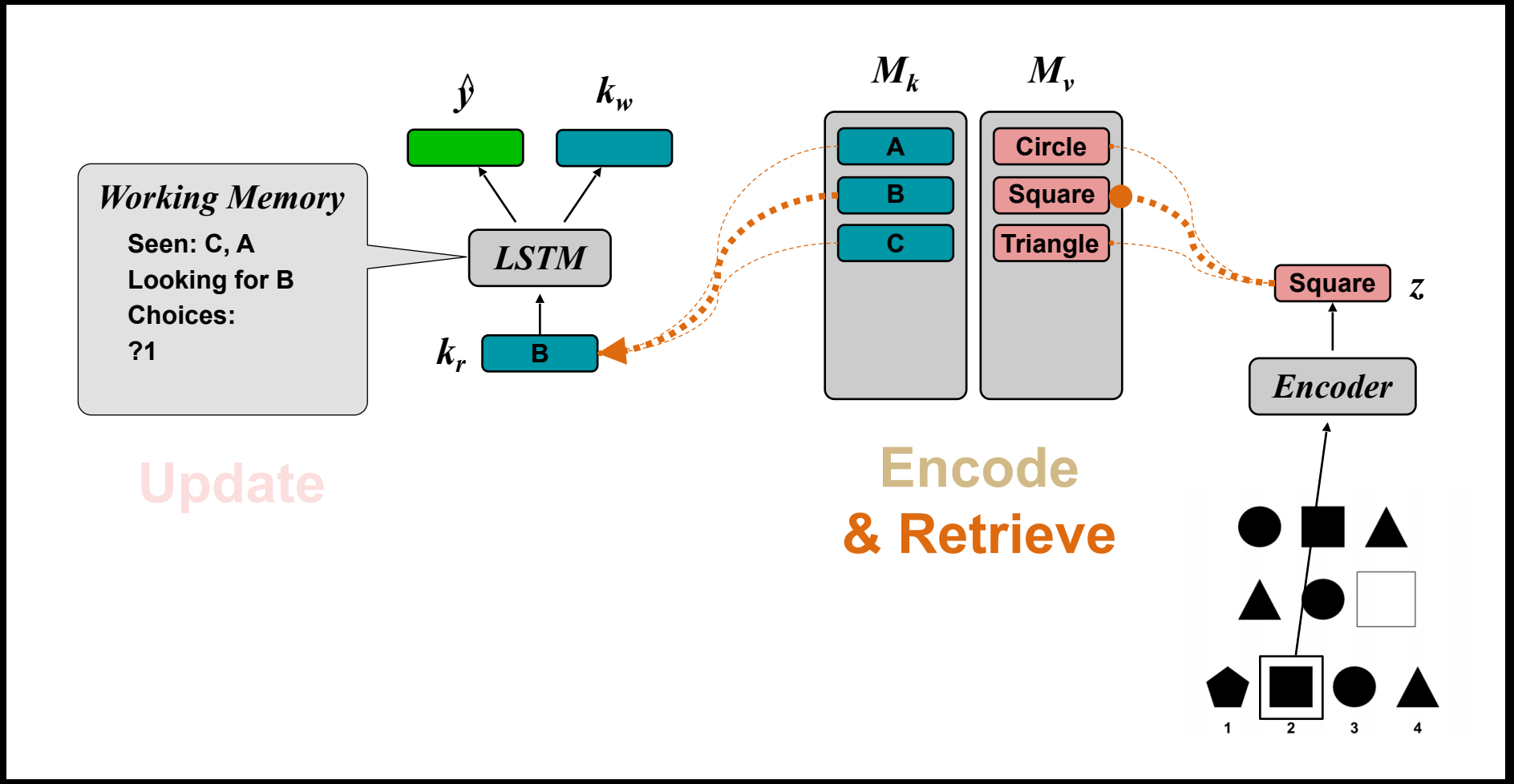
Example Trial: Distribution of Three



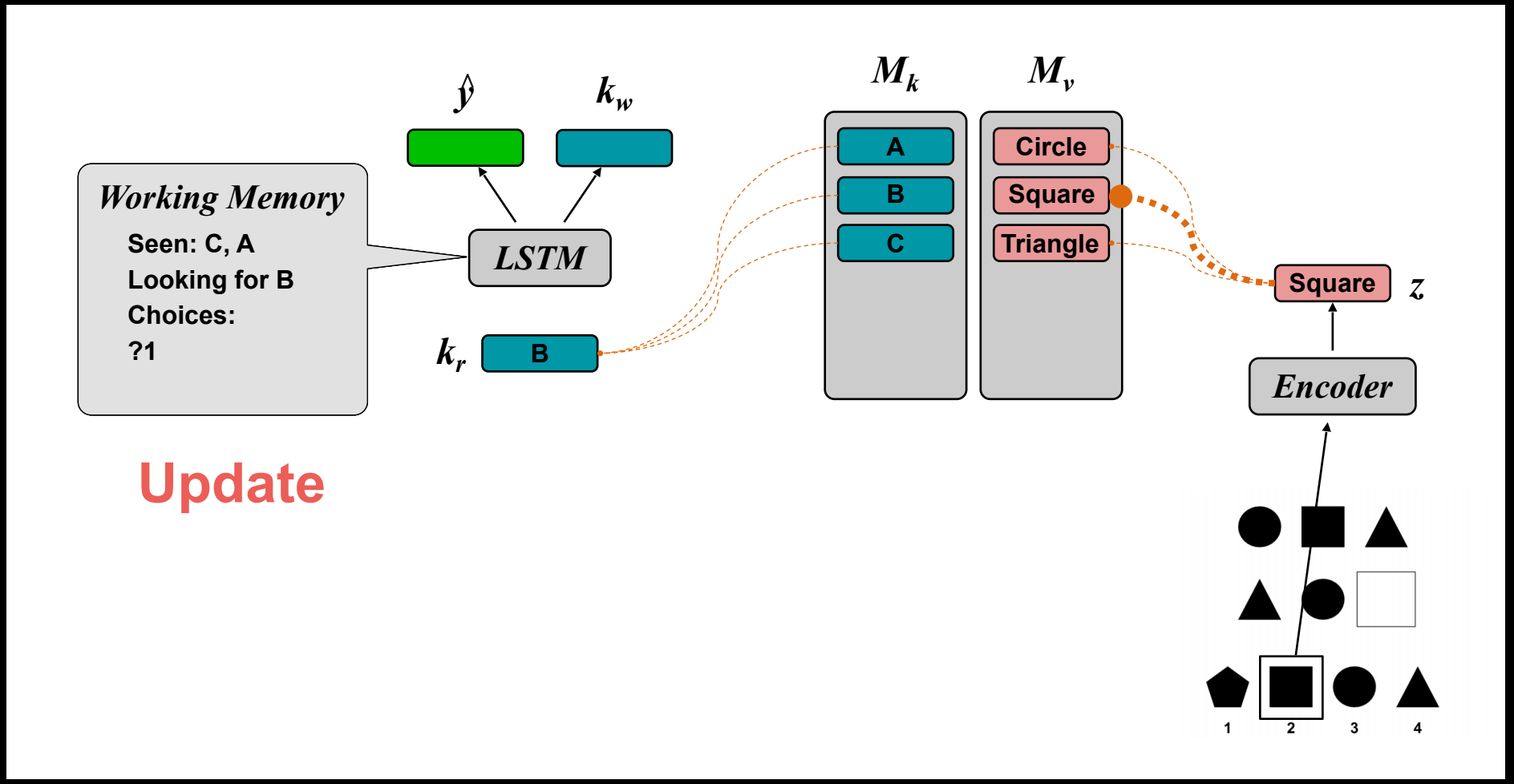
Example Trial: Distribution of Three



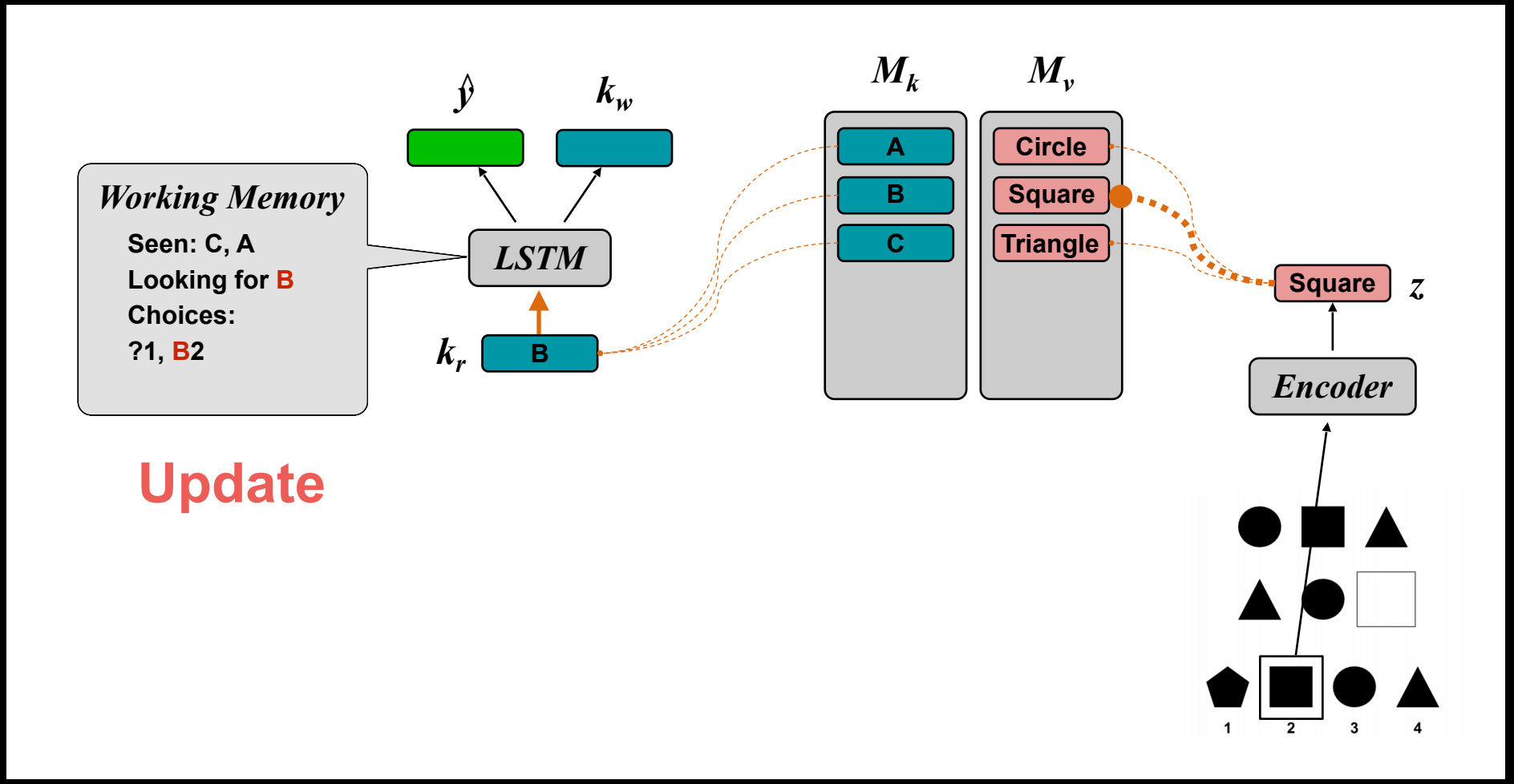
Example Trial: Distribution of Three



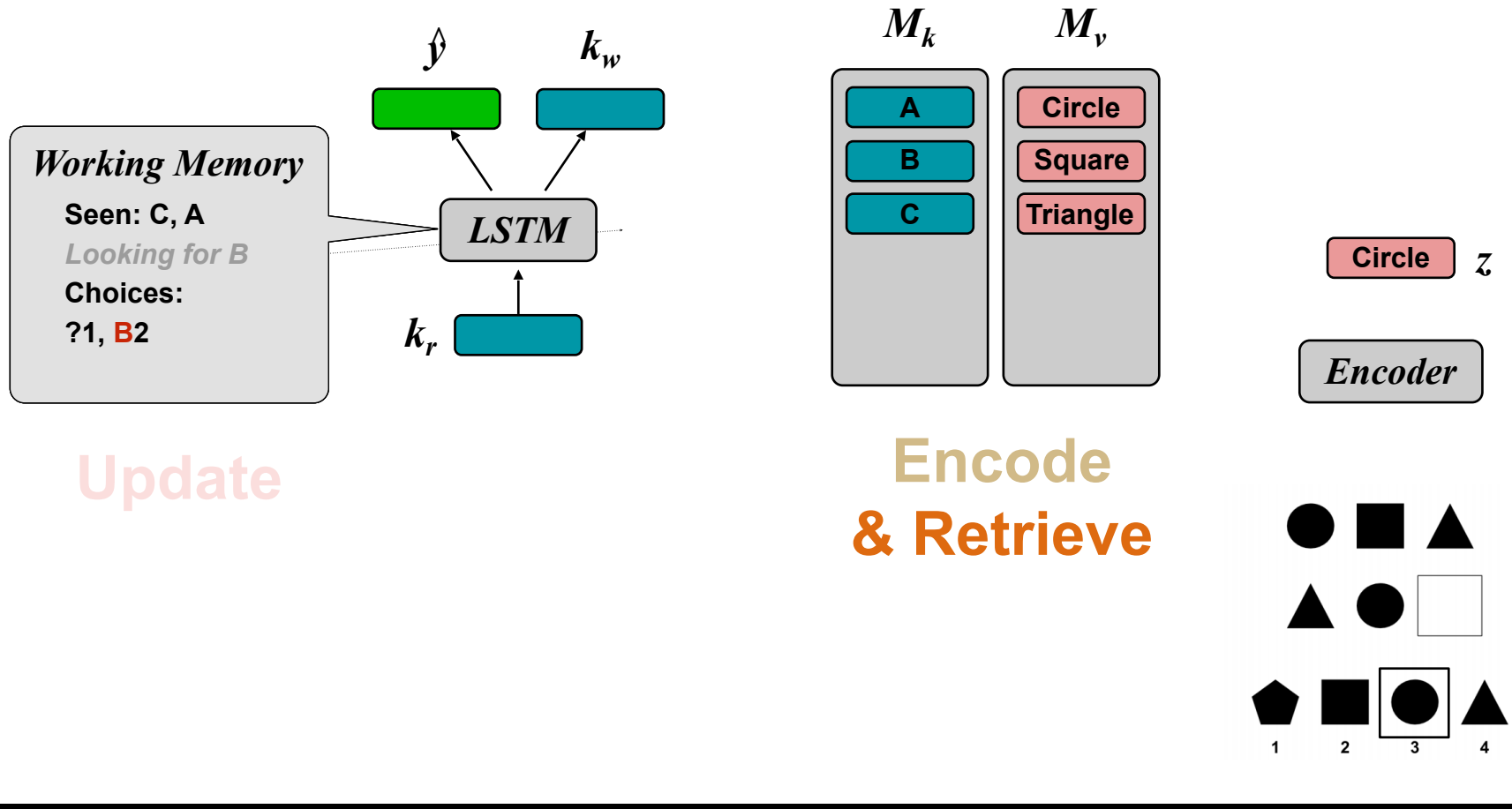
Example Trial: Distribution of Three



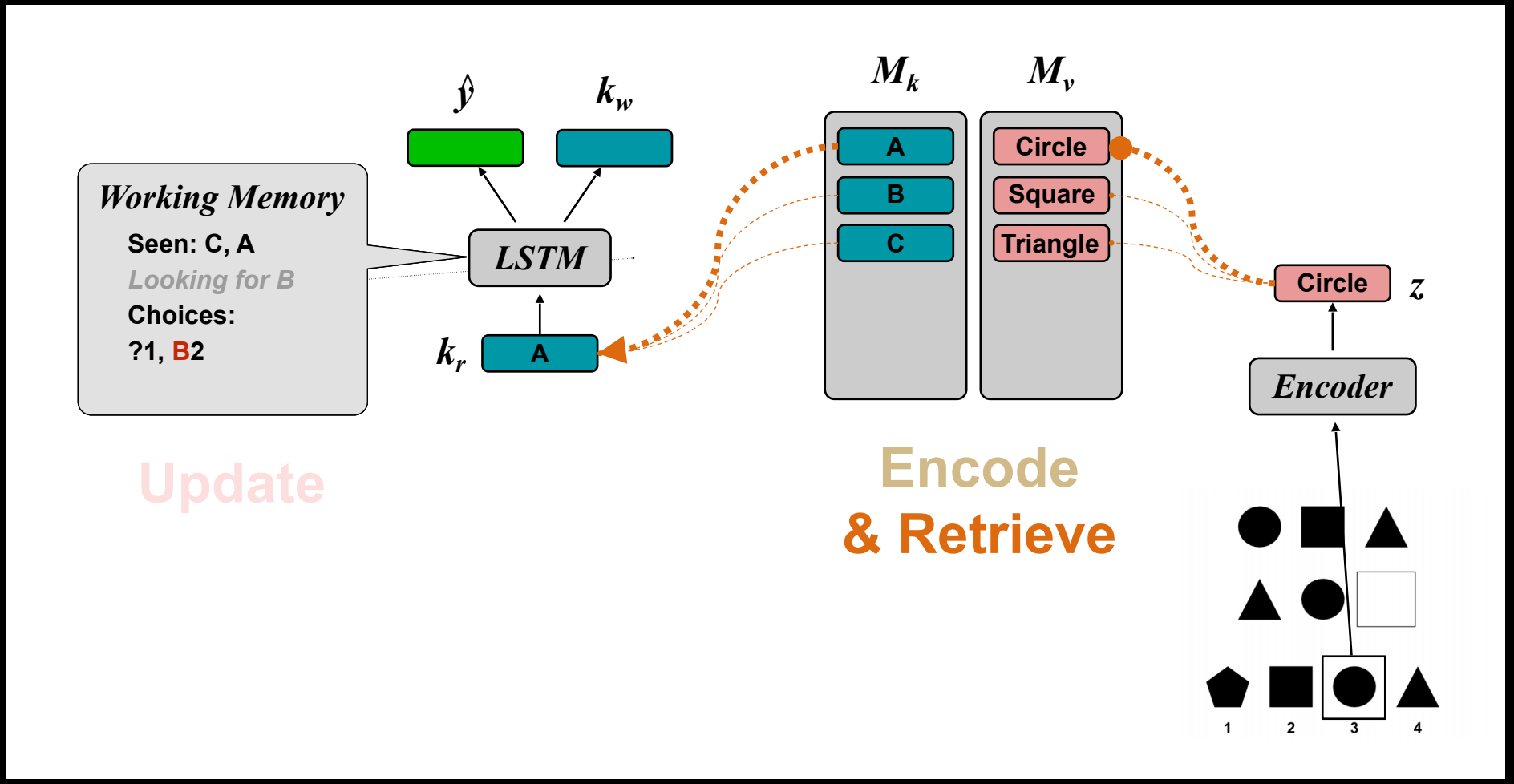
Example Trial: Distribution of Three



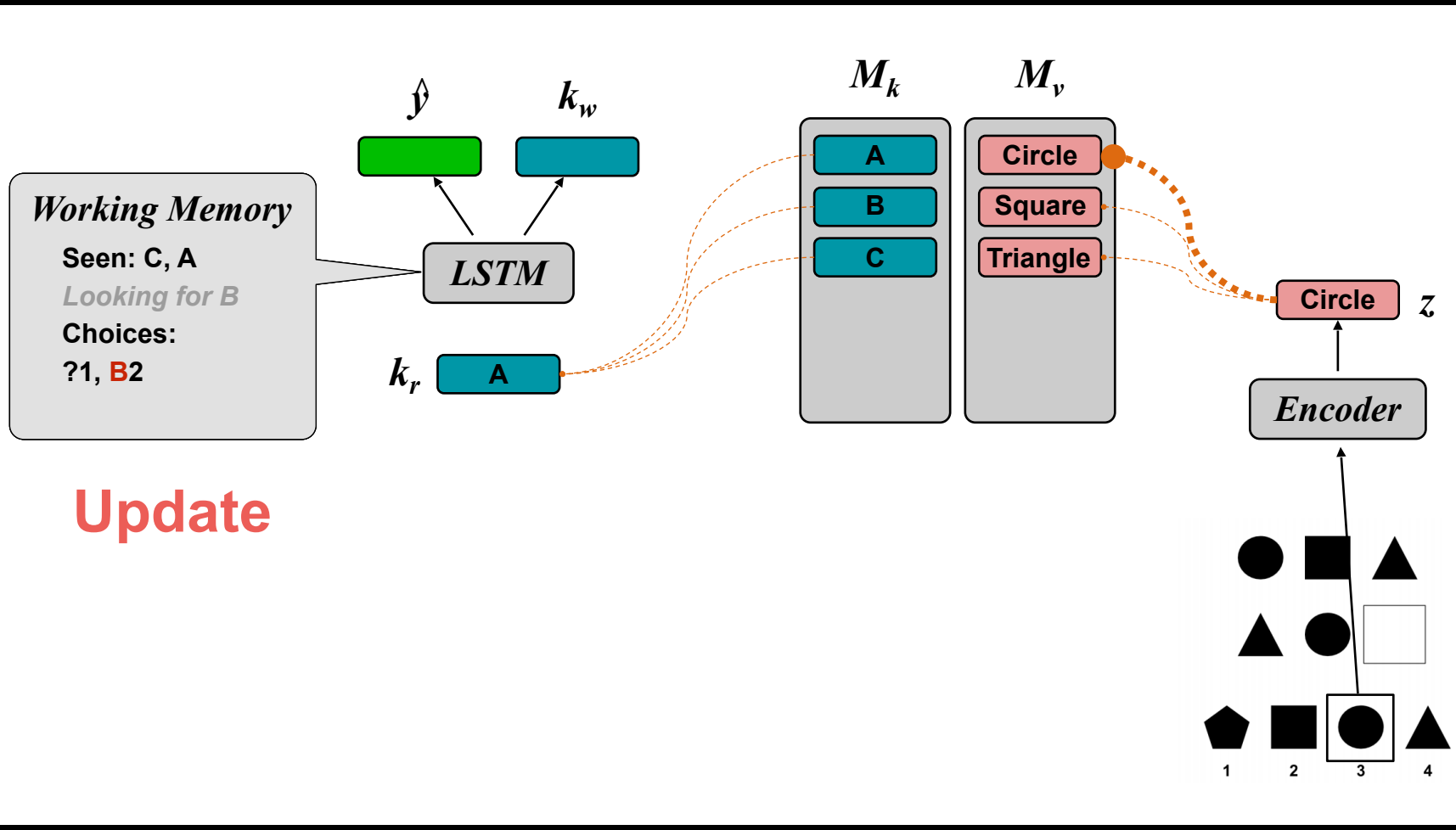
Example Trial: Distribution of Three



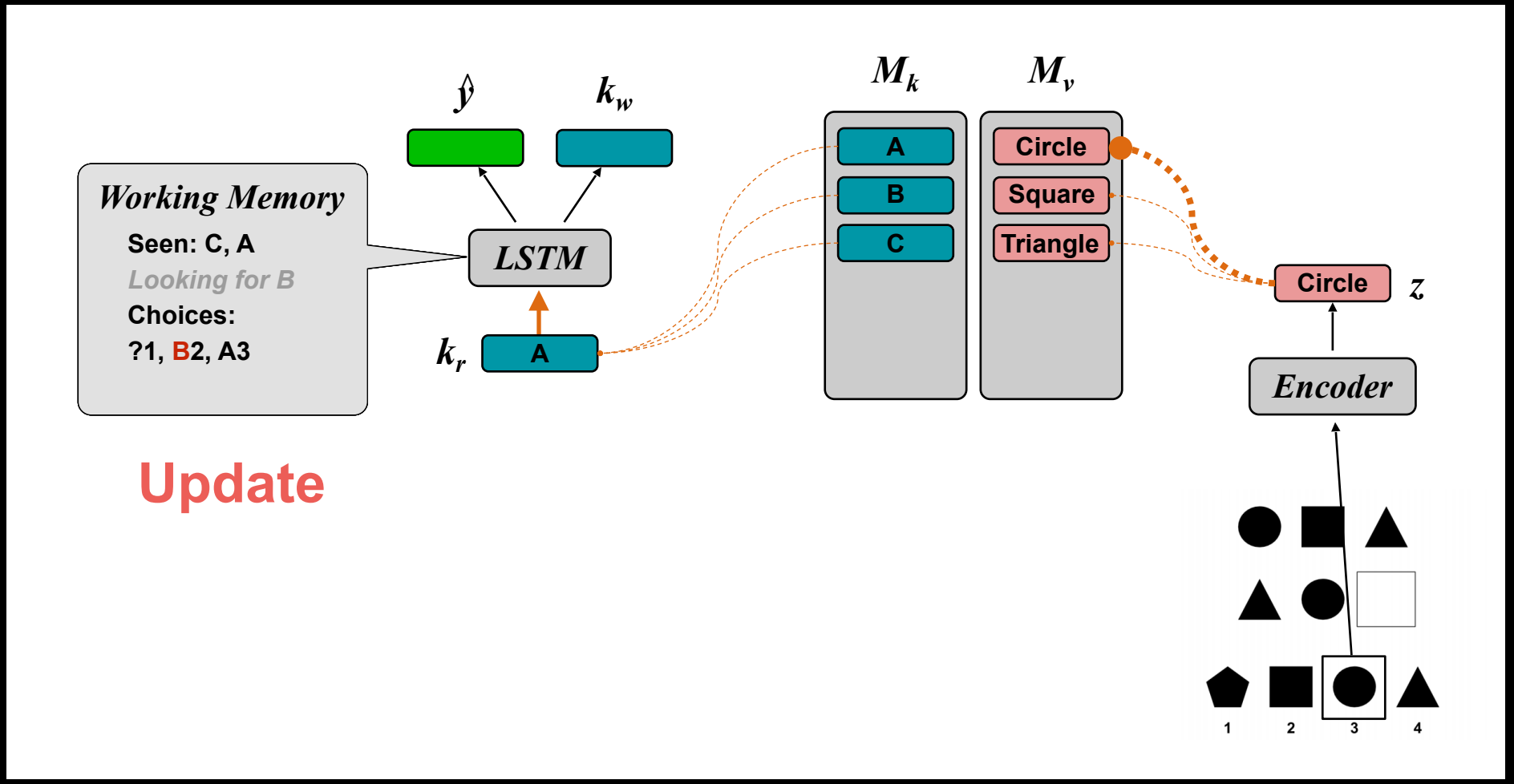
Example Trial: Distribution of Three



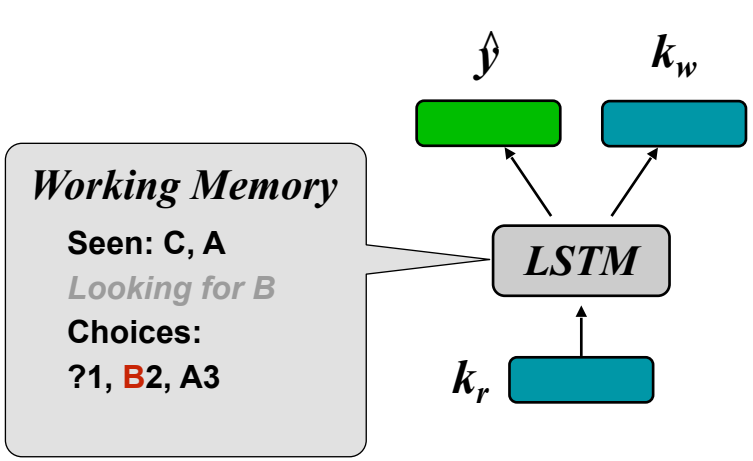
Example Trial: Distribution of Three



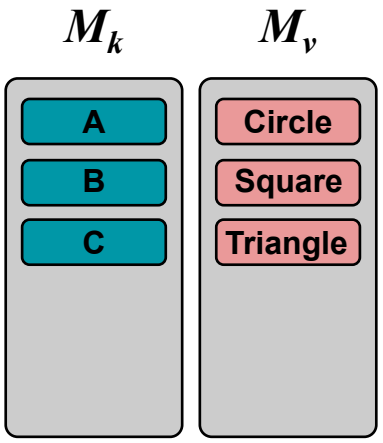
Example Trial: Distribution of Three



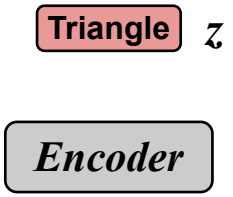
Example Trial: Distribution of Three

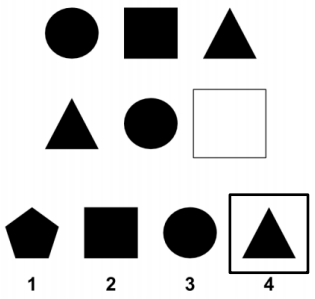


Update

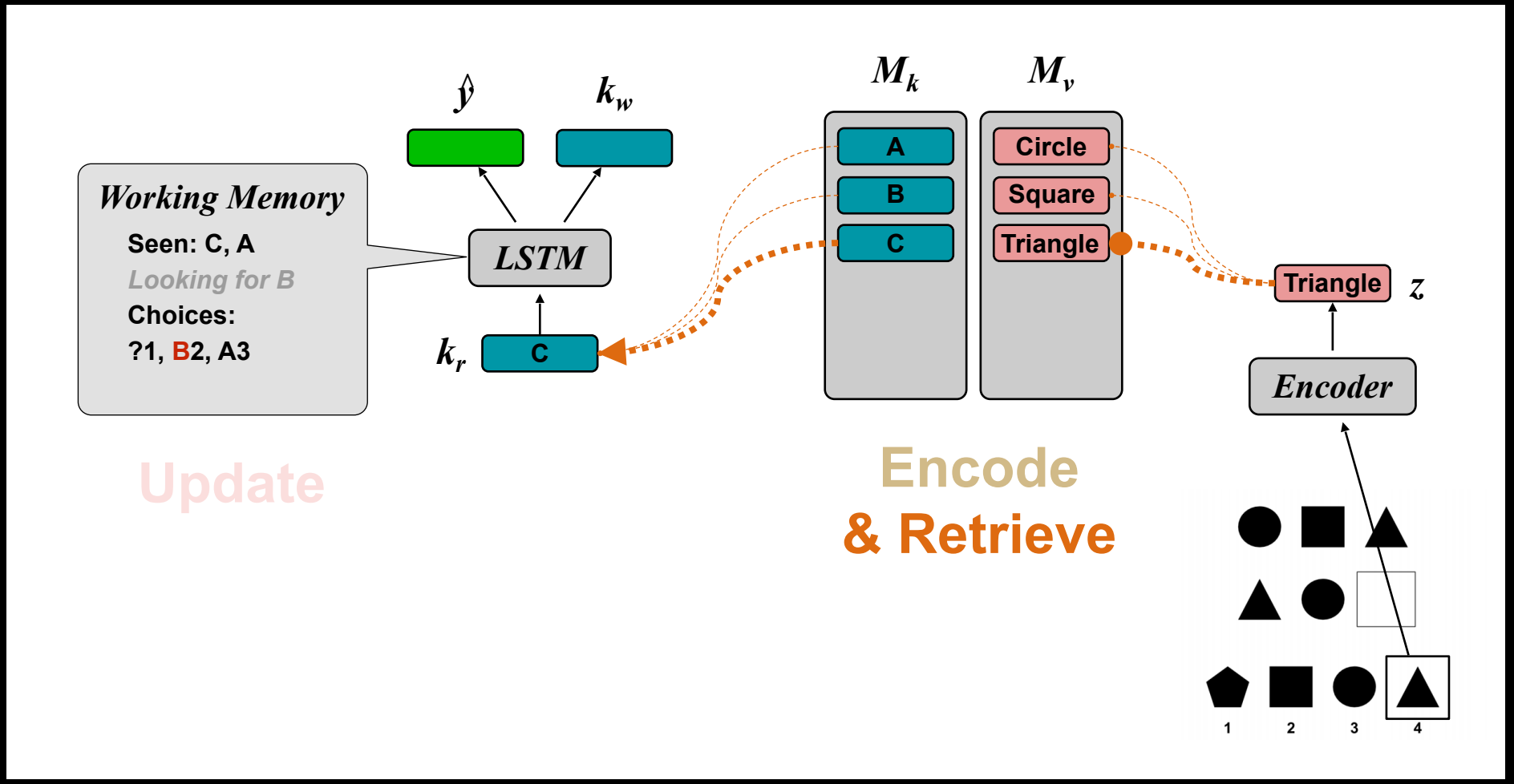


Encode
& Retrieve

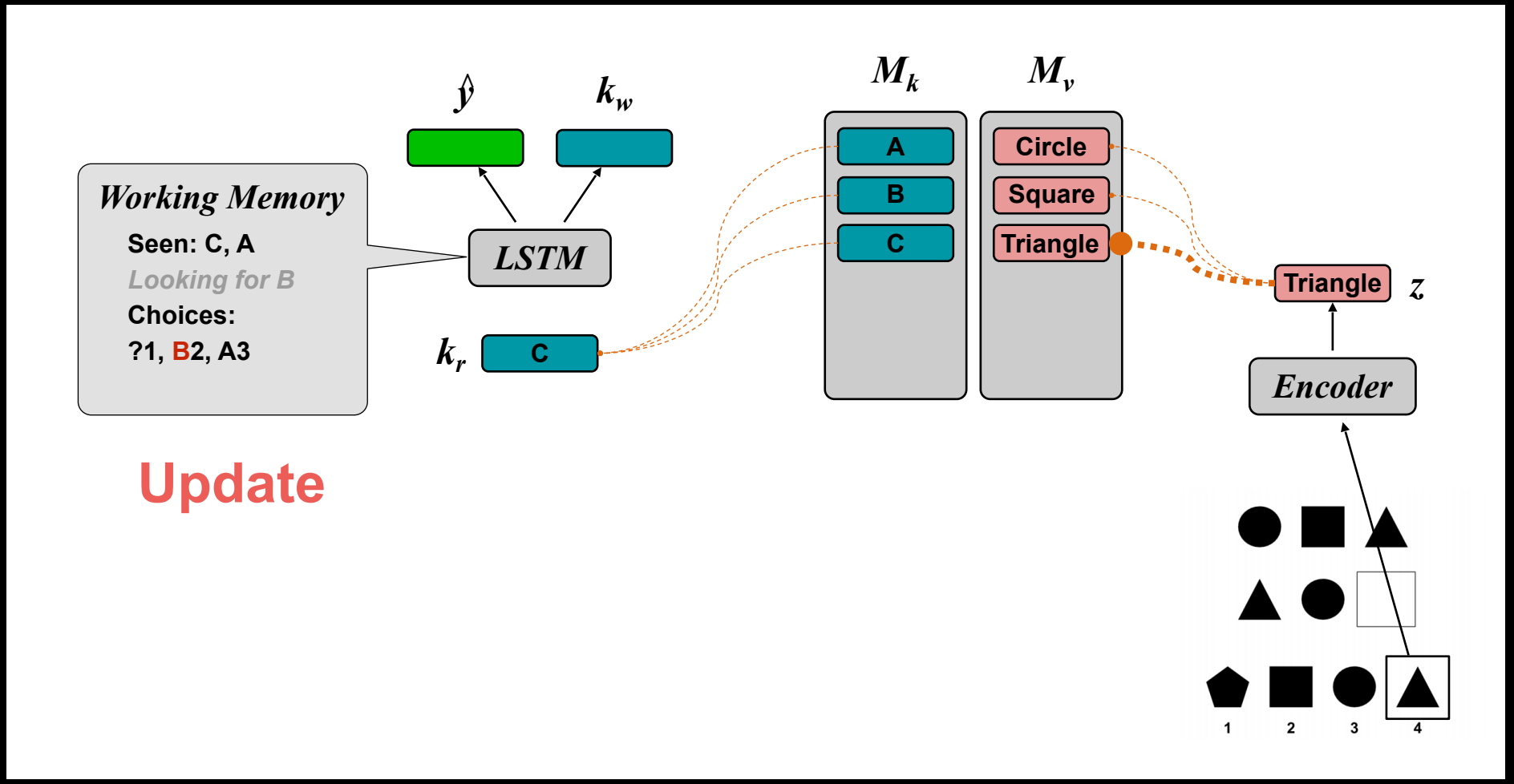




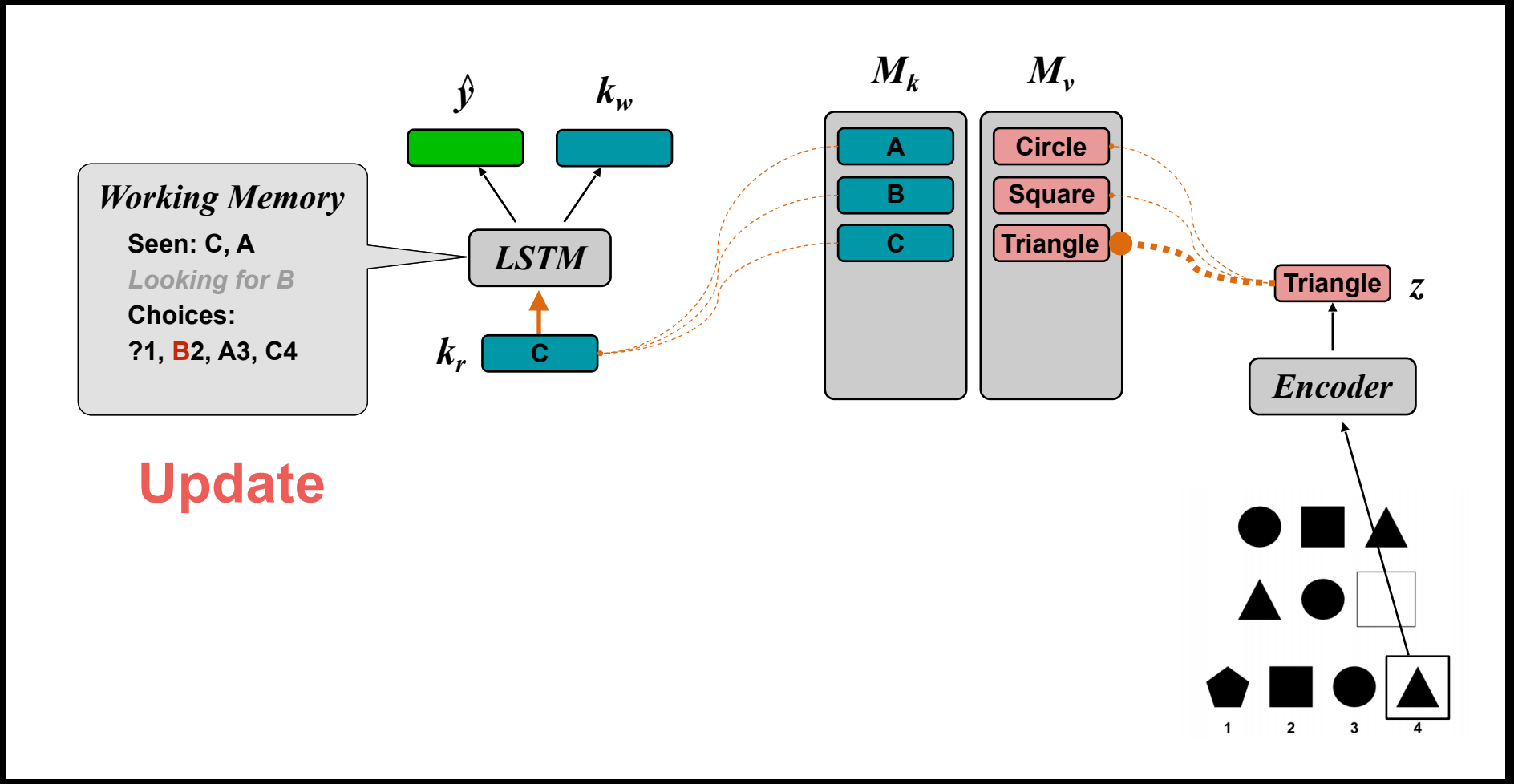
Example Trial: Distribution of Three



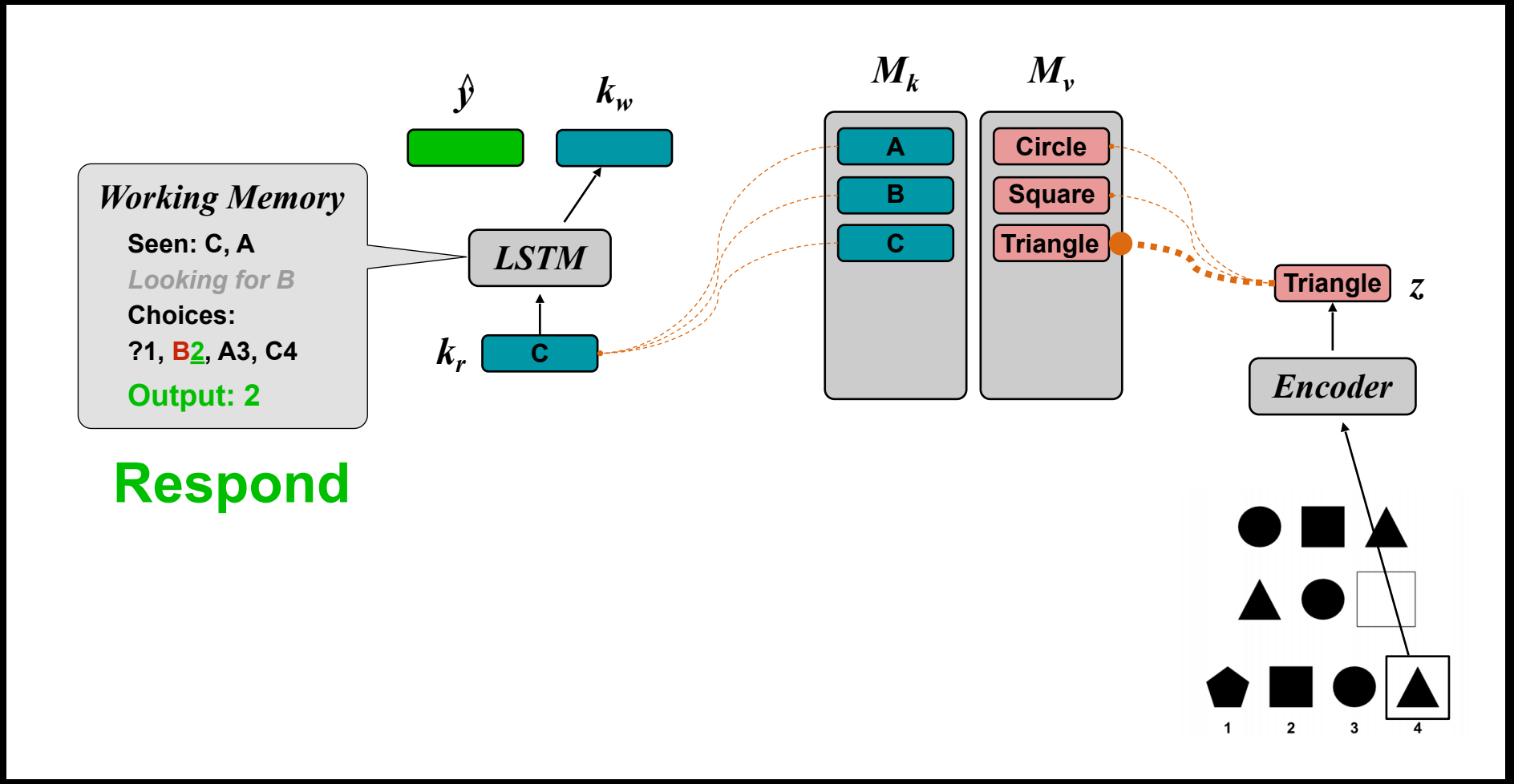
Example Trial: Distribution of Three



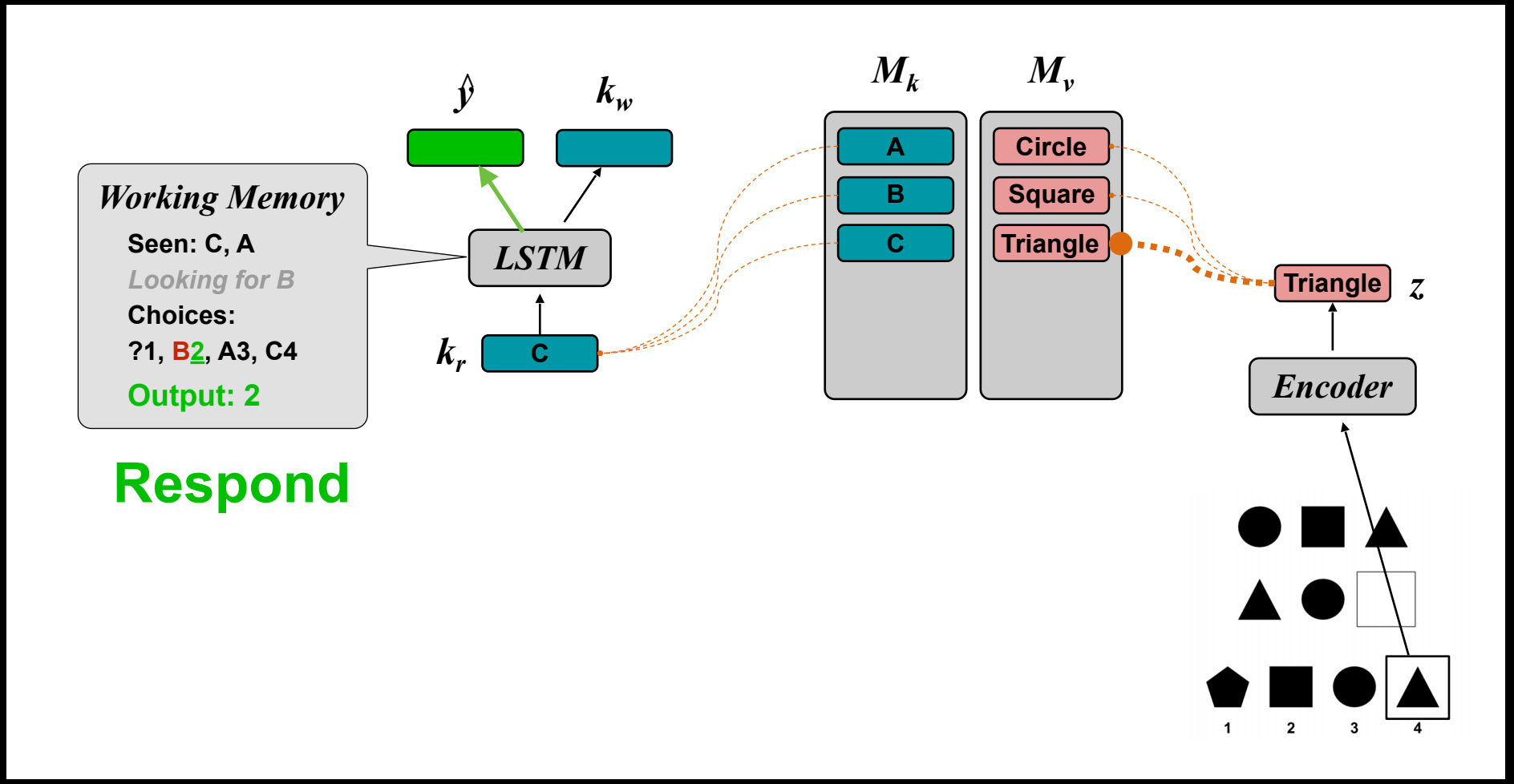
Example Trial: Distribution of Three



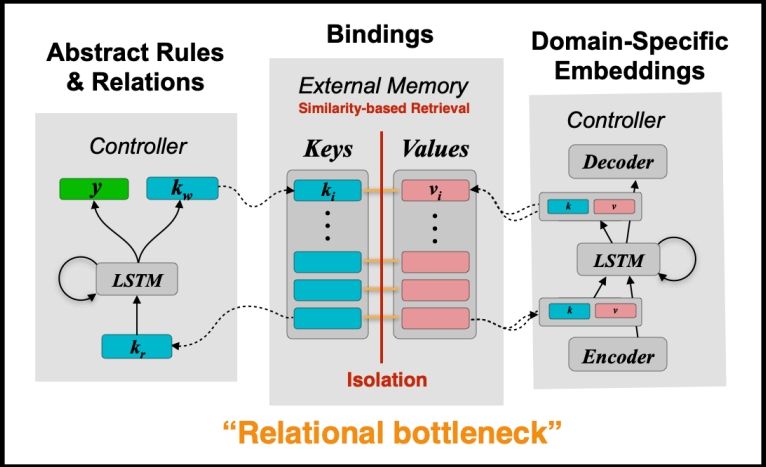
Example Trial: Distribution of Three



Example Trial: Distribution of Three



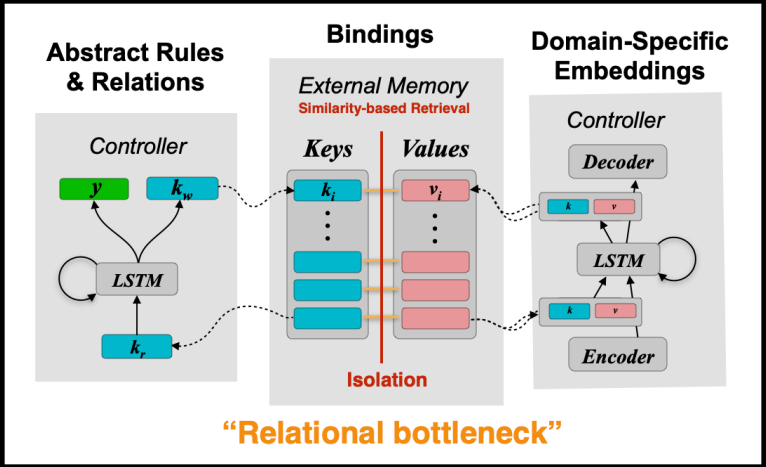
Neural Network with External Memory



Neural Network with External Memory

External memory for isolation + similarity-based retrieval

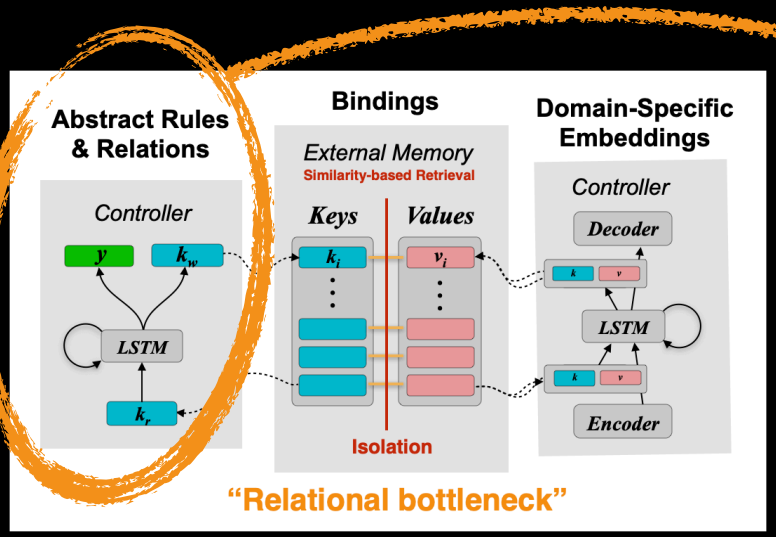
⇒ relational bottleneck



Neural Network with External Memory

External memory for isolation + similarity-based retrieval

⇒ relational bottleneck



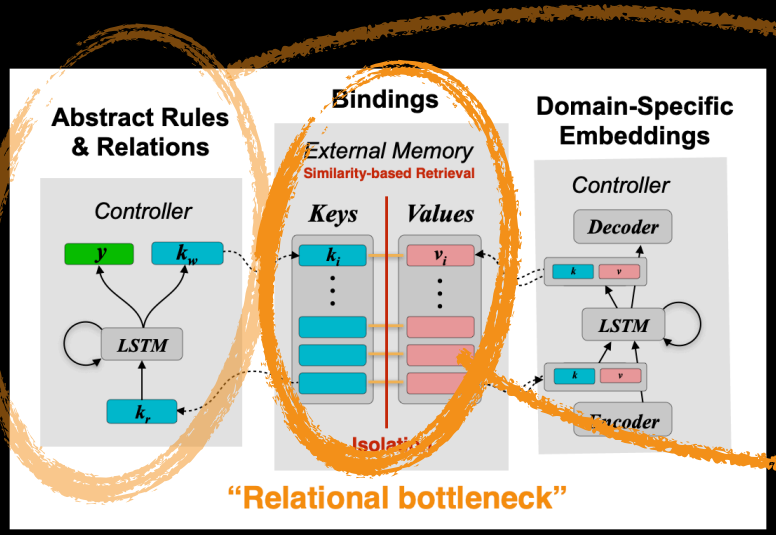
$$v = f(x, y, z)$$

⇒ ***learning*** of abstract functions

Neural Network with External Memory

External memory for isolation + similarity-based retrieval

⇒ relational bottleneck



$$v = f(x, y, z)$$

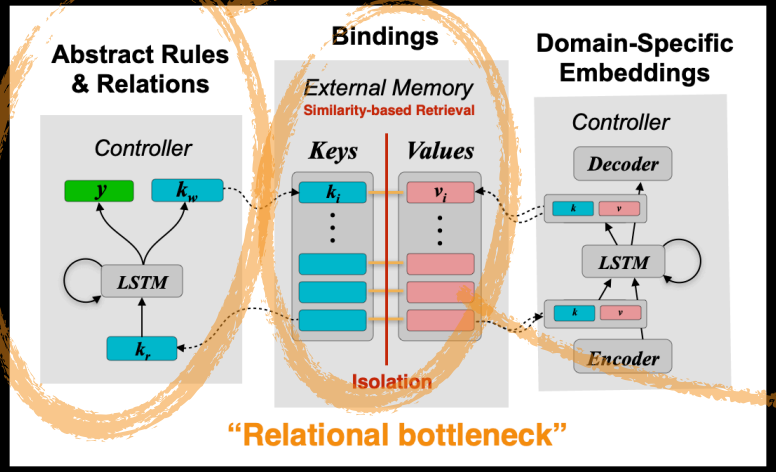
⇒ *learning* of abstract functions

+ variable binding

Neural Network with External Memory

External memory for isolation + similarity-based retrieval

⇒ relational bottleneck



$$v = f(x, y, z)$$

⇒ *learning* of abstract functions

+ variable binding

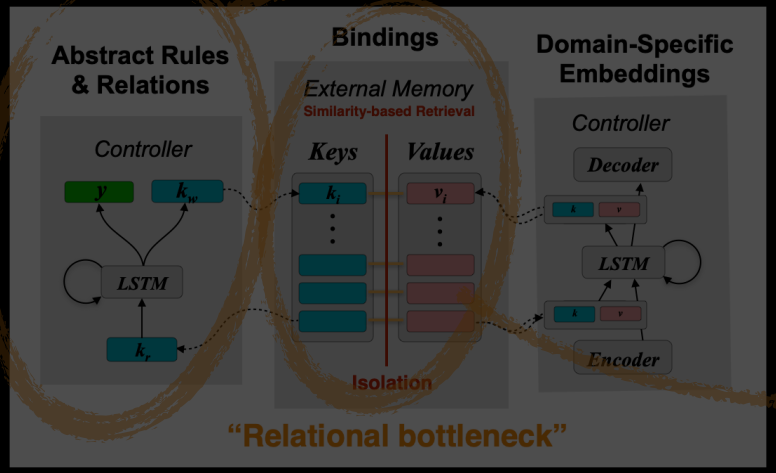
⇒ emergence of

symbolic computation

Neural Network with External Memory

External memory for isolation + similarity-based retrieval

⇒ relational bottleneck



$$v = f(x, y, z)$$

⇒ *learning* of abstract functions

+ variable binding

⇒ emergence of *symbolic computation*

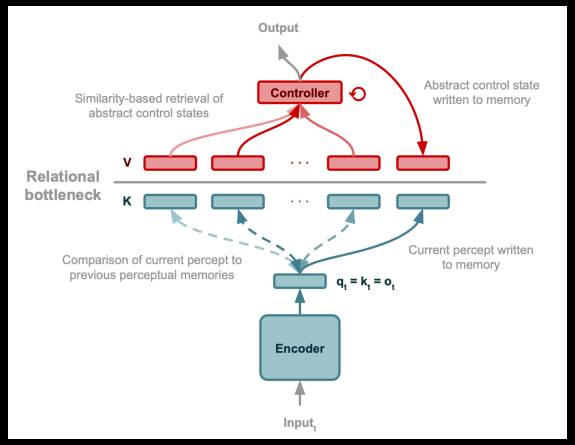
- Can be implemented in a variety of forms…

Relational Bottleneck

(Webb et al., 2023)

LSTM + External Memory

ESBN *(Webb, Sinha & Cohen, 2021)*

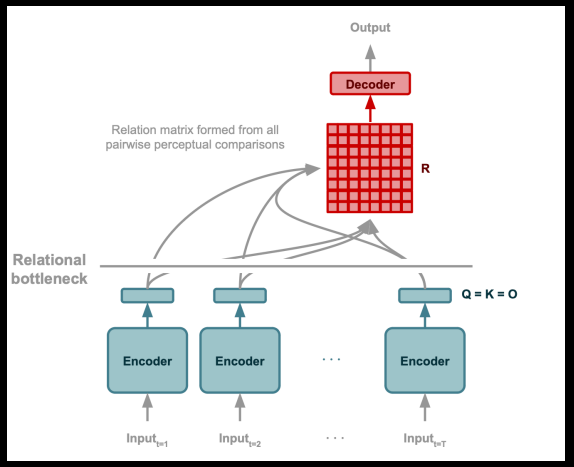


Relational Bottleneck

(Webb et al., 2023)

Simple Contrastive Networks

CorelNet *(Kerg et al., 2022)*

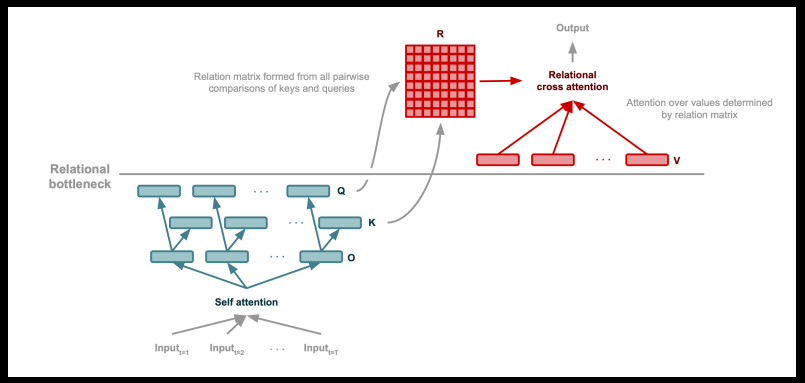


Relational Bottleneck

(Webb et al., 2023)

Transformers (with Relational Cross-Attention)

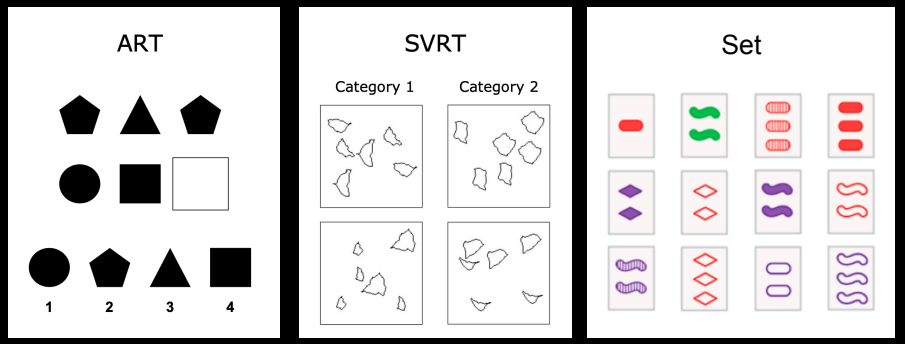
Abstractor *(Altabaa, Webb, Cohen & Lafferty, 2023)*



Relational Bottleneck

Visual Reasoning

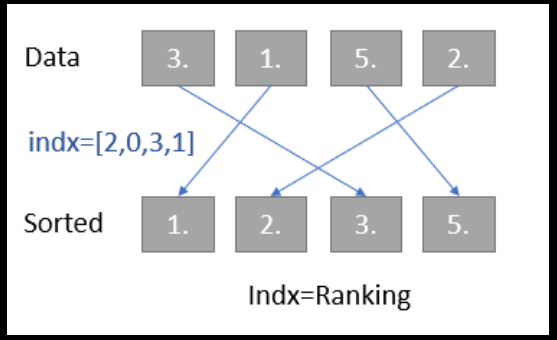
Webb, Mondal & Cohen (NeurIPS 2024)



Relational Bottleneck

Sorting

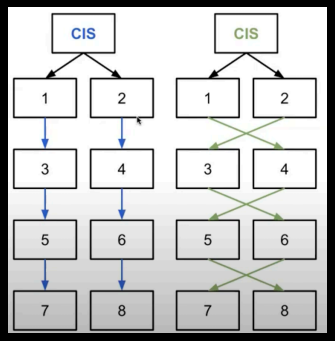
(Altabaa, Webb, Cohen & Lafferty (ICLR 2024)

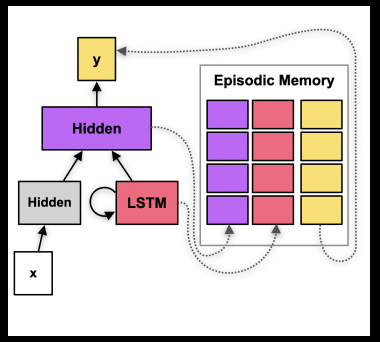


Relational Bottleneck

Sequence Learning

Giallanza, Campbell & Cohen (OpenMind, 2024)



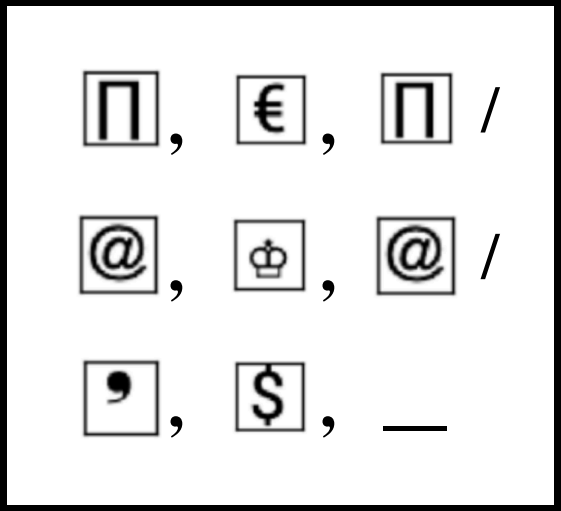


Symbolic Attention Heads in LLMs

(Yang et al., 2025)

Abstract Sequence Task

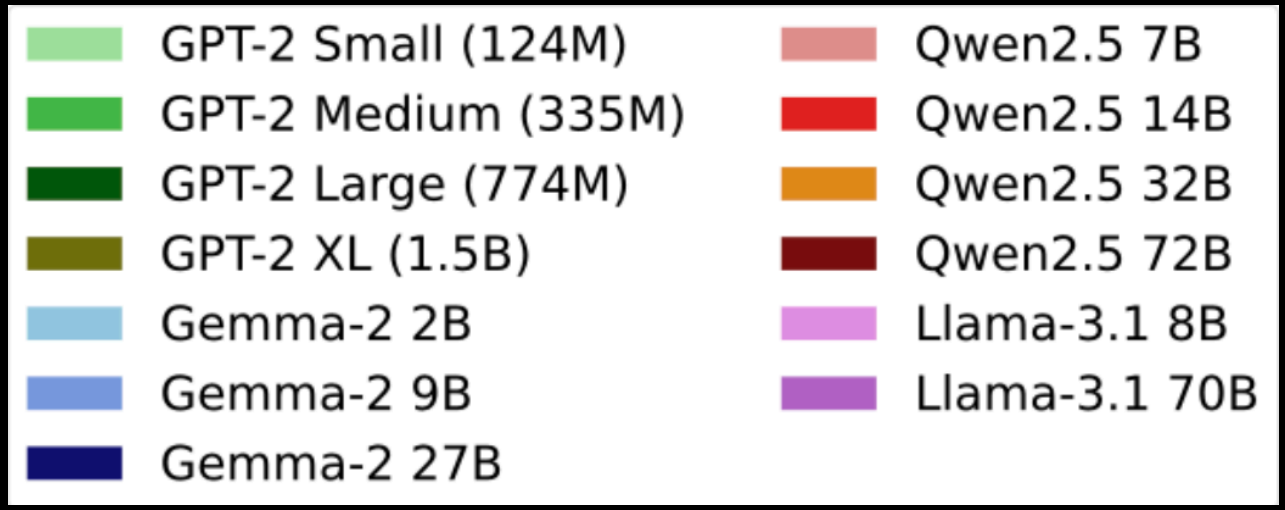
("A-B-A", "A-B-B"…)



Symbolic Attention Heads in LLMs

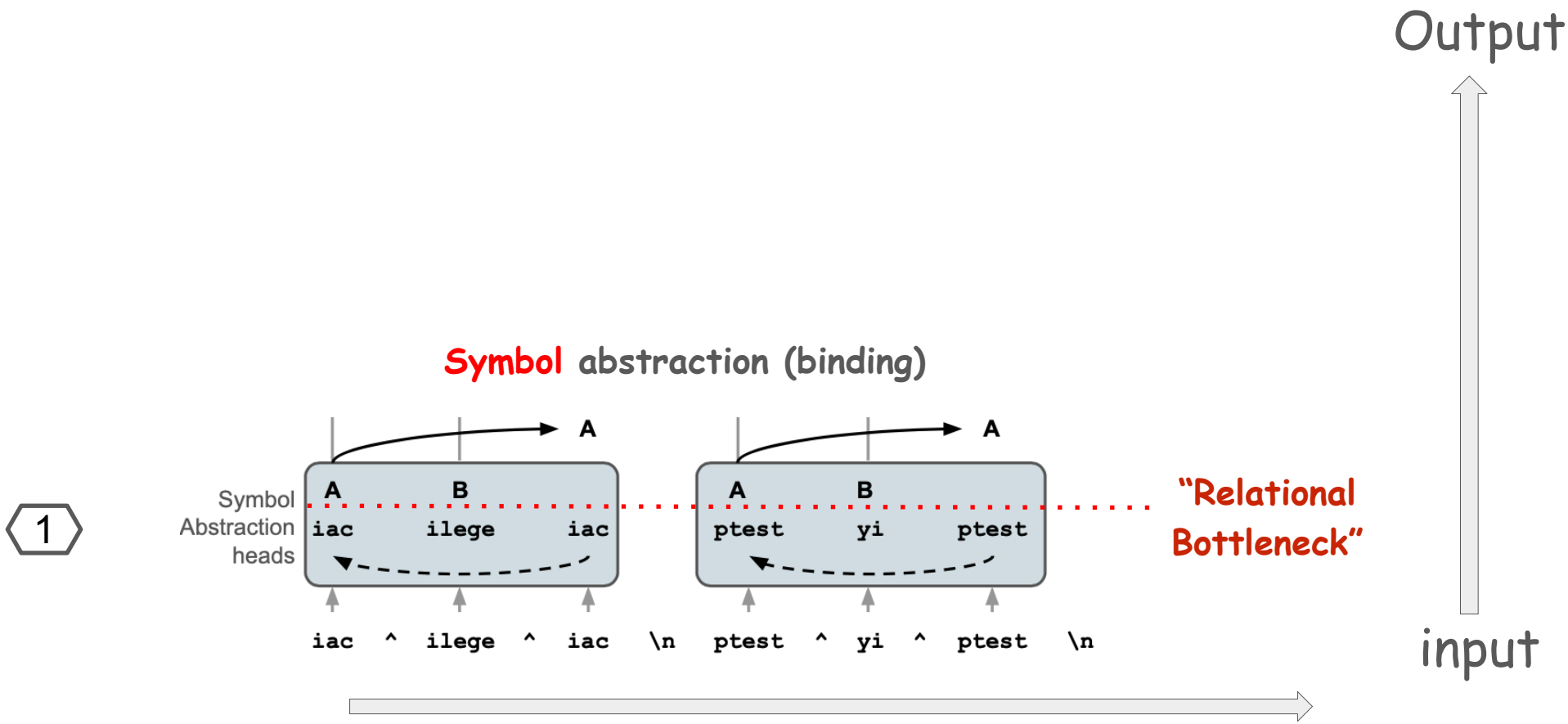
(Yang et al., 2025)

LLM's
(w/ access to internal reps)



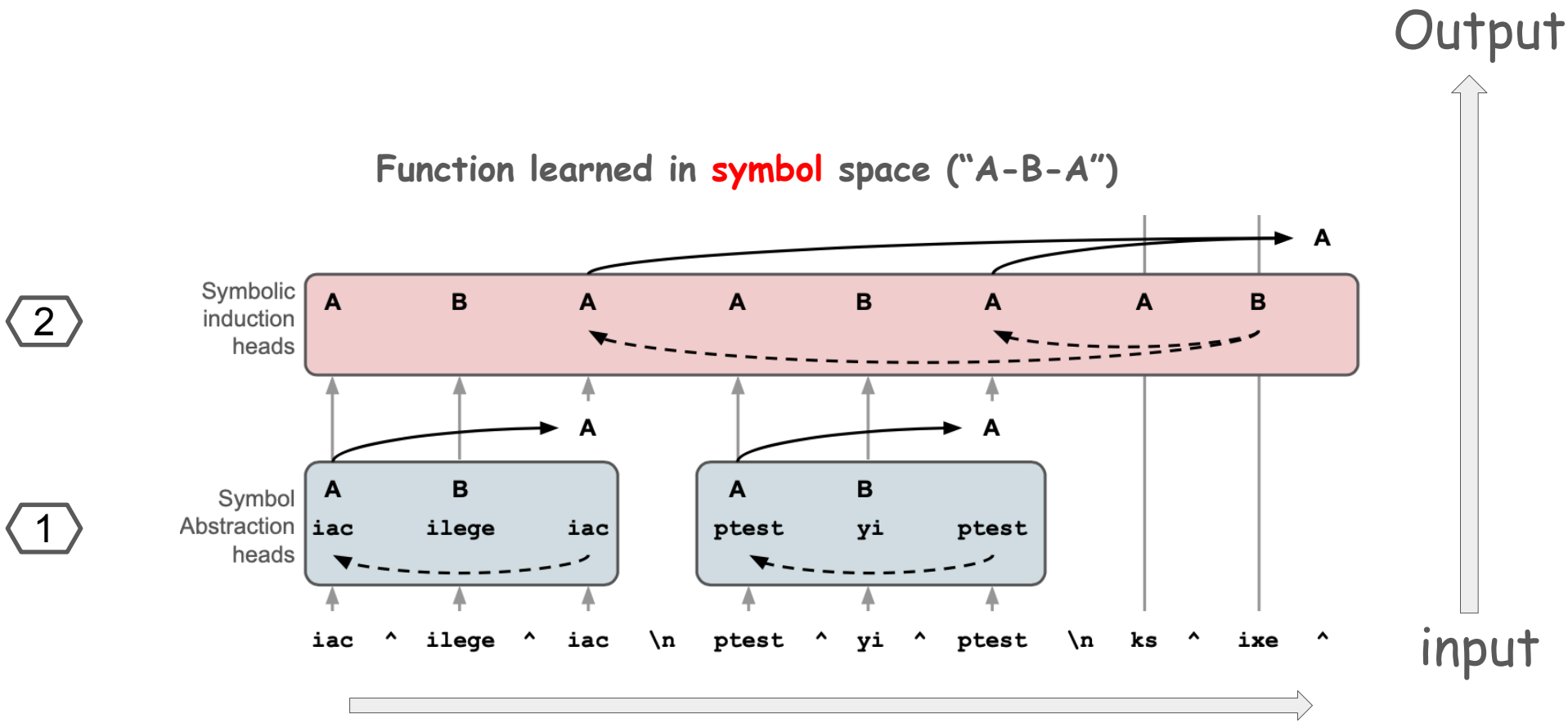
Emergent Symbolic Processing Mechanism in LLMs

(Yang et al., ICLR 2025)



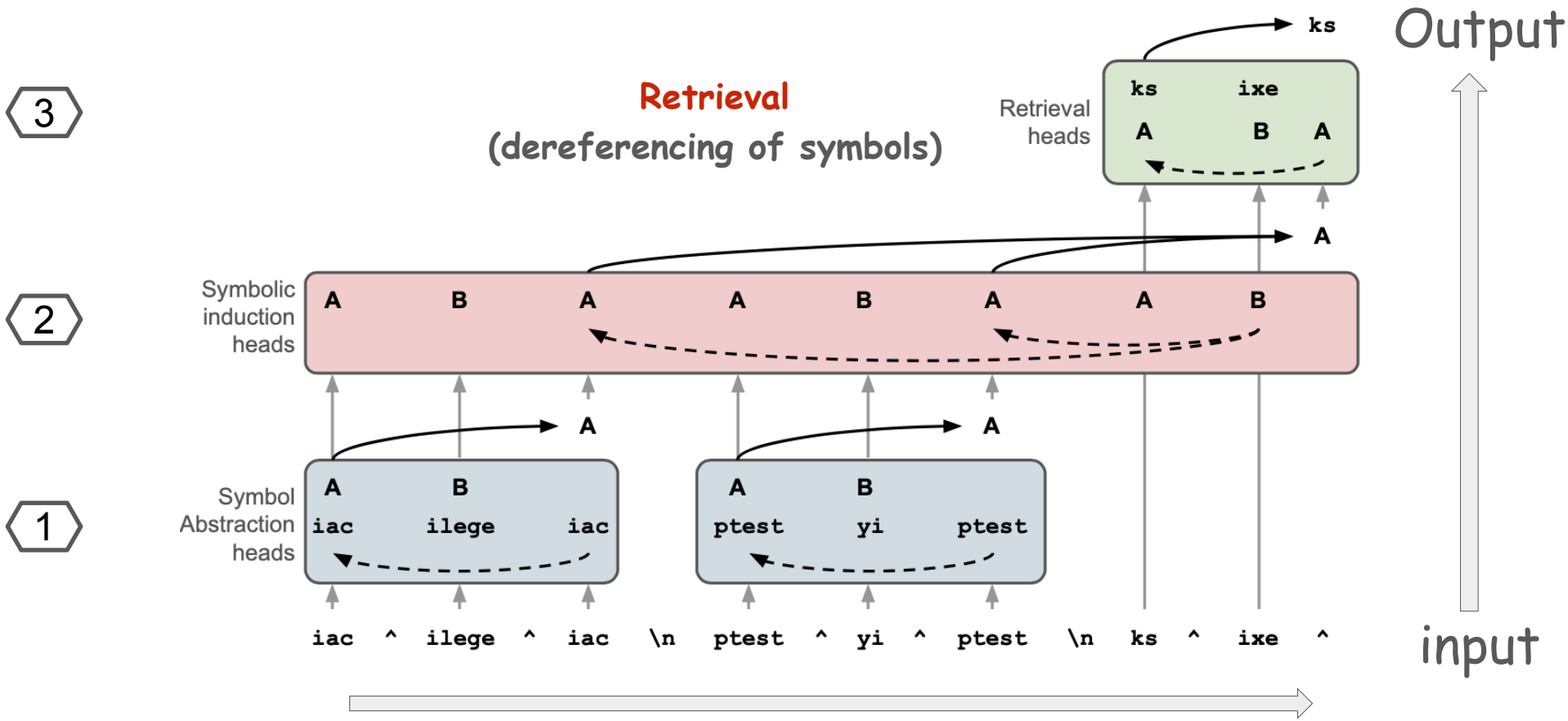
Emergent Symbolic Processing Mechanism in LLMs

(Yang et al., ICLR 2025)



Emergent Symbolic Processing Mechanism in LLMs

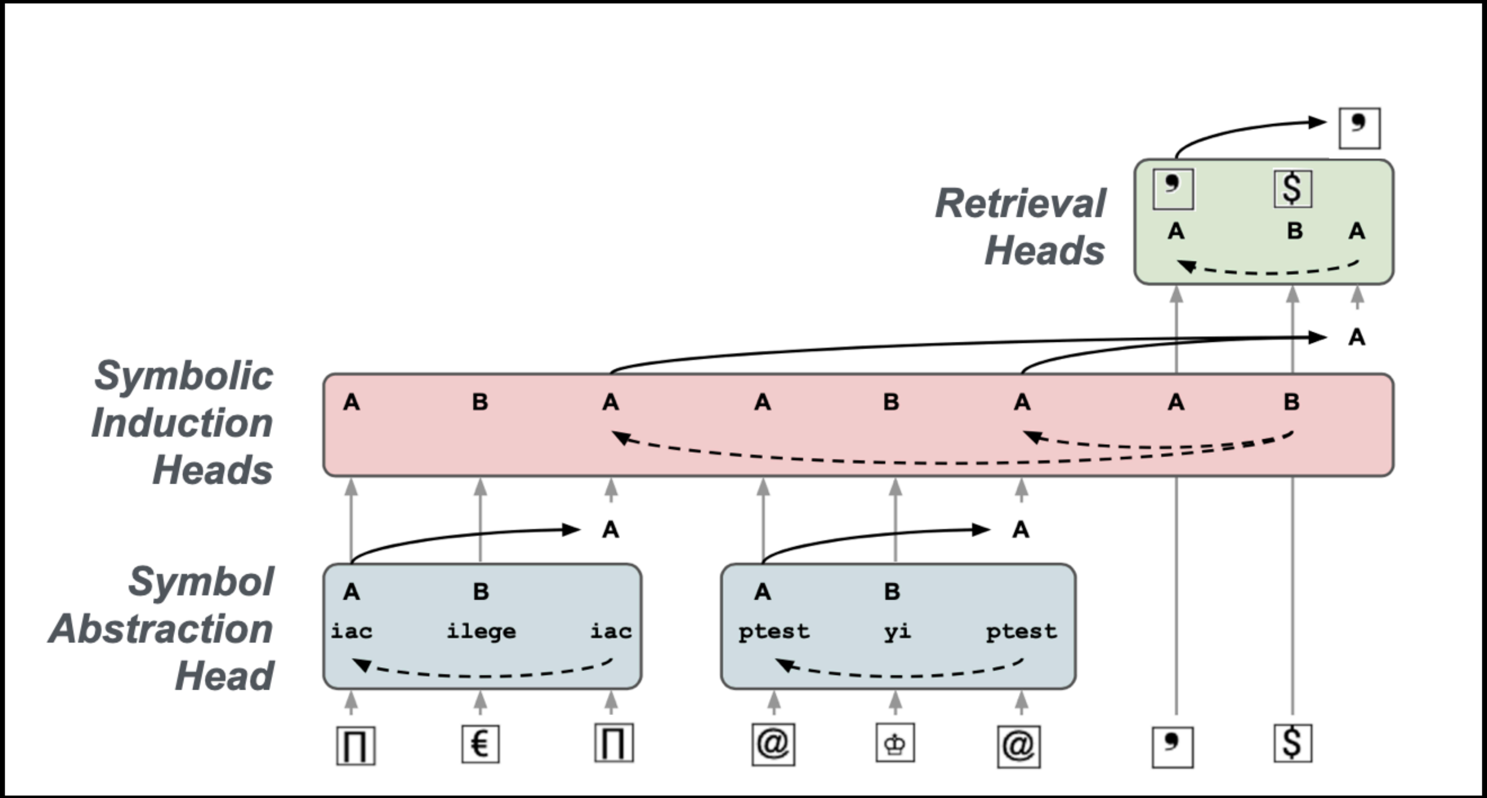
(Yang et al., ICLR 2025)



Symbolic Attention Heads in LLMs

(Yang et al., 2025)

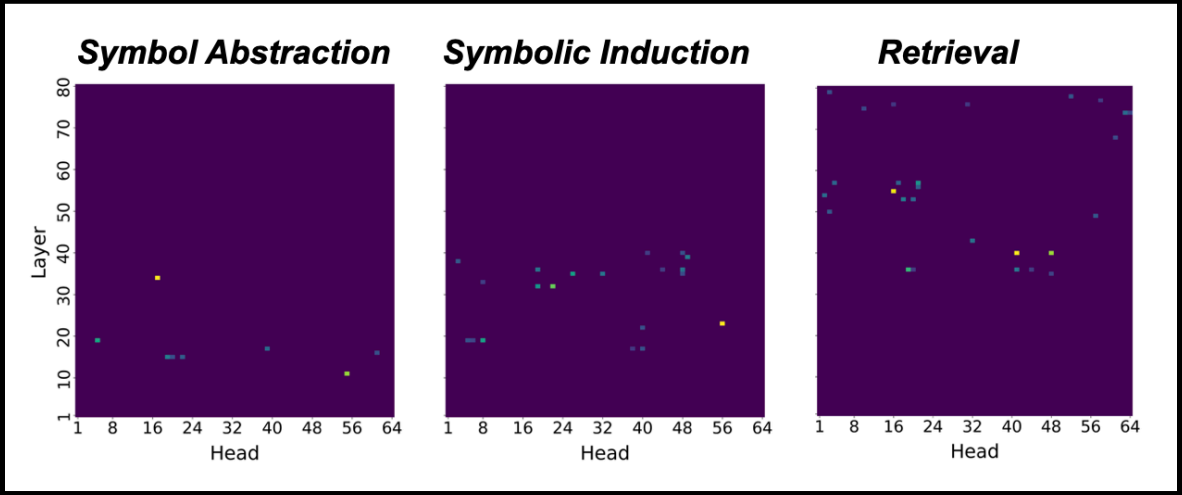
Hypothesized Functions



Symbolic Attention Heads in LLMs

(Yang et al., 2025)

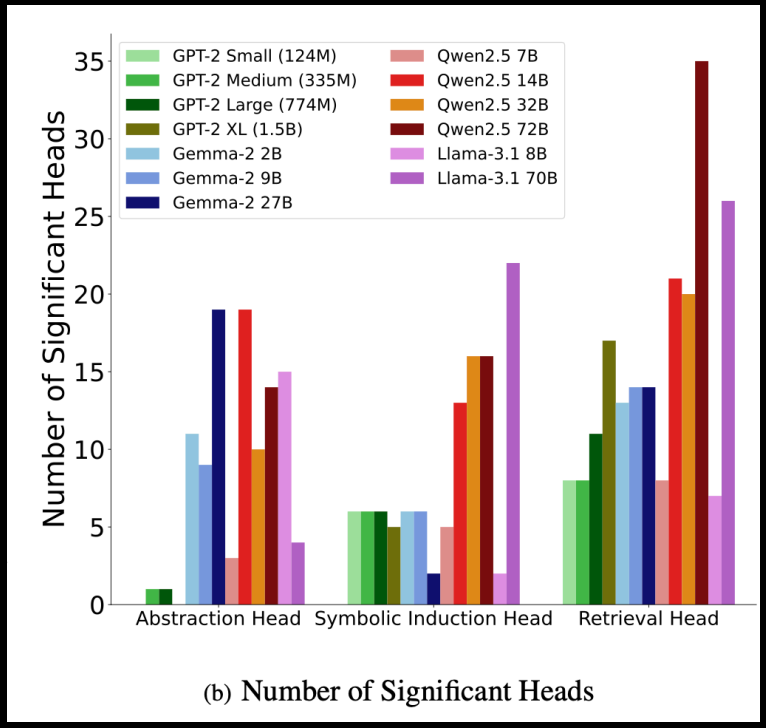
Causal Mediation Analysis



Symbolic Attention Heads in LLMs

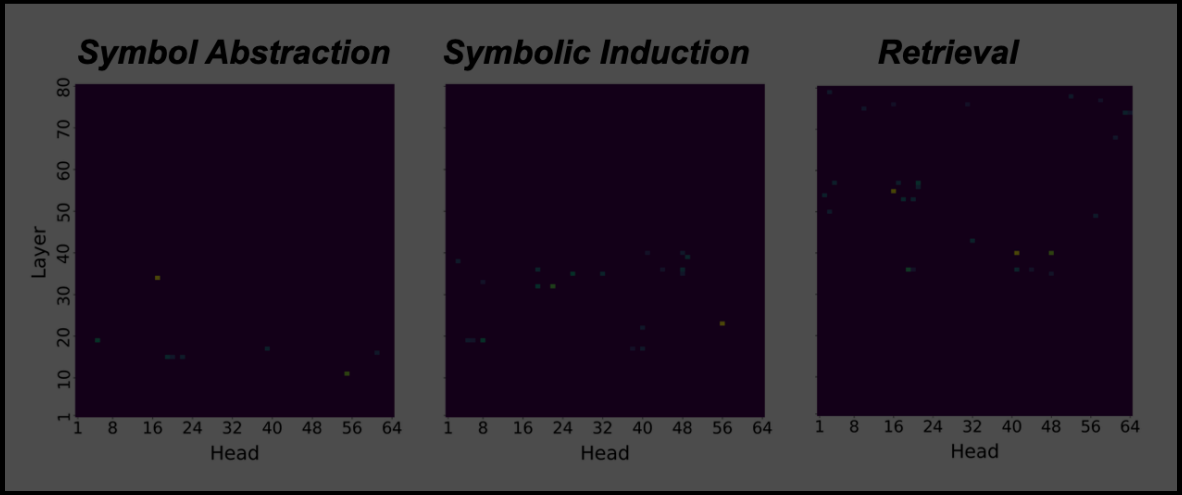
(Yang et al., 2025)

Observed across foundation models



(b) Number of Significant Heads

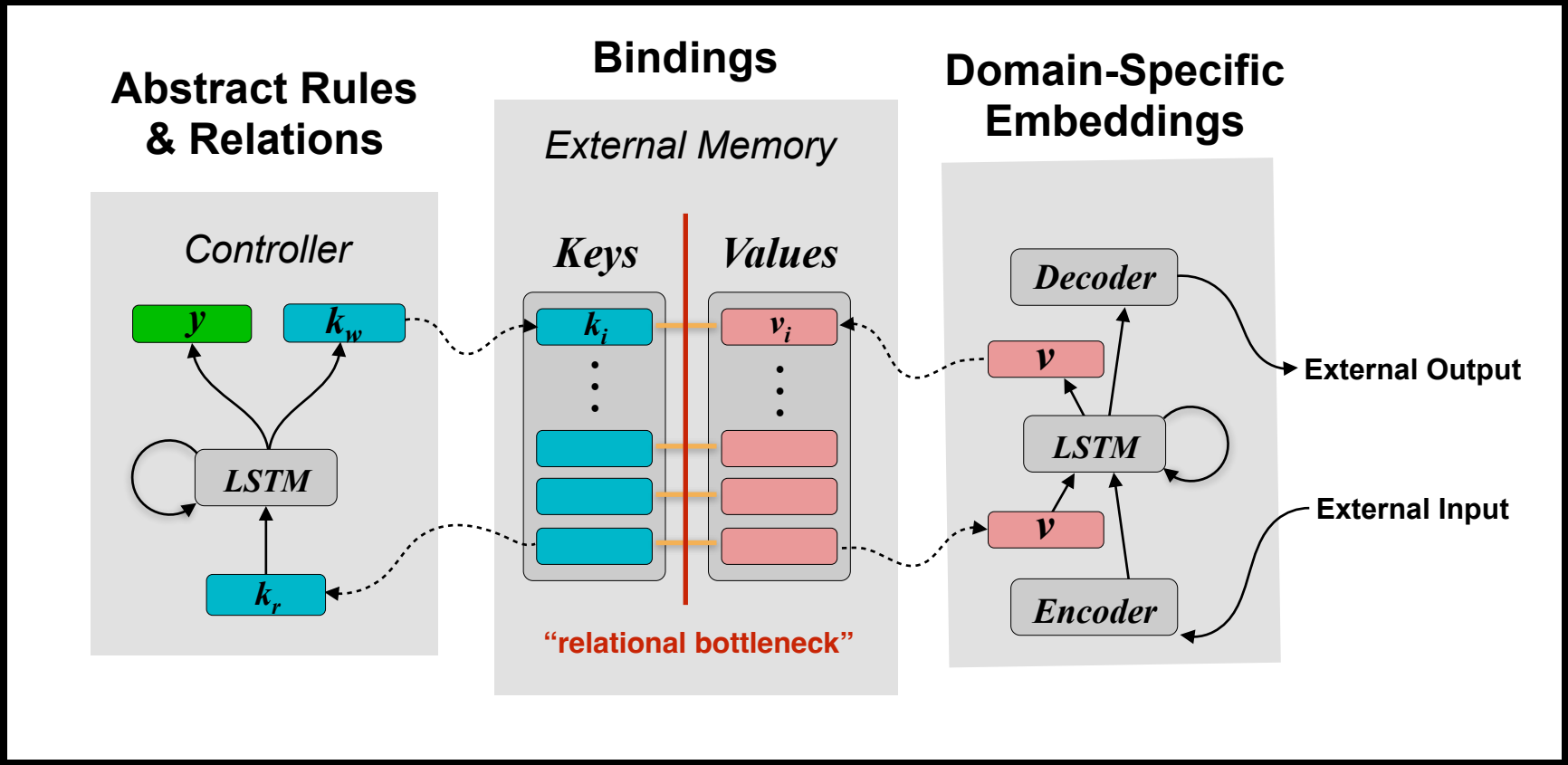
Causal Mediation Analysis



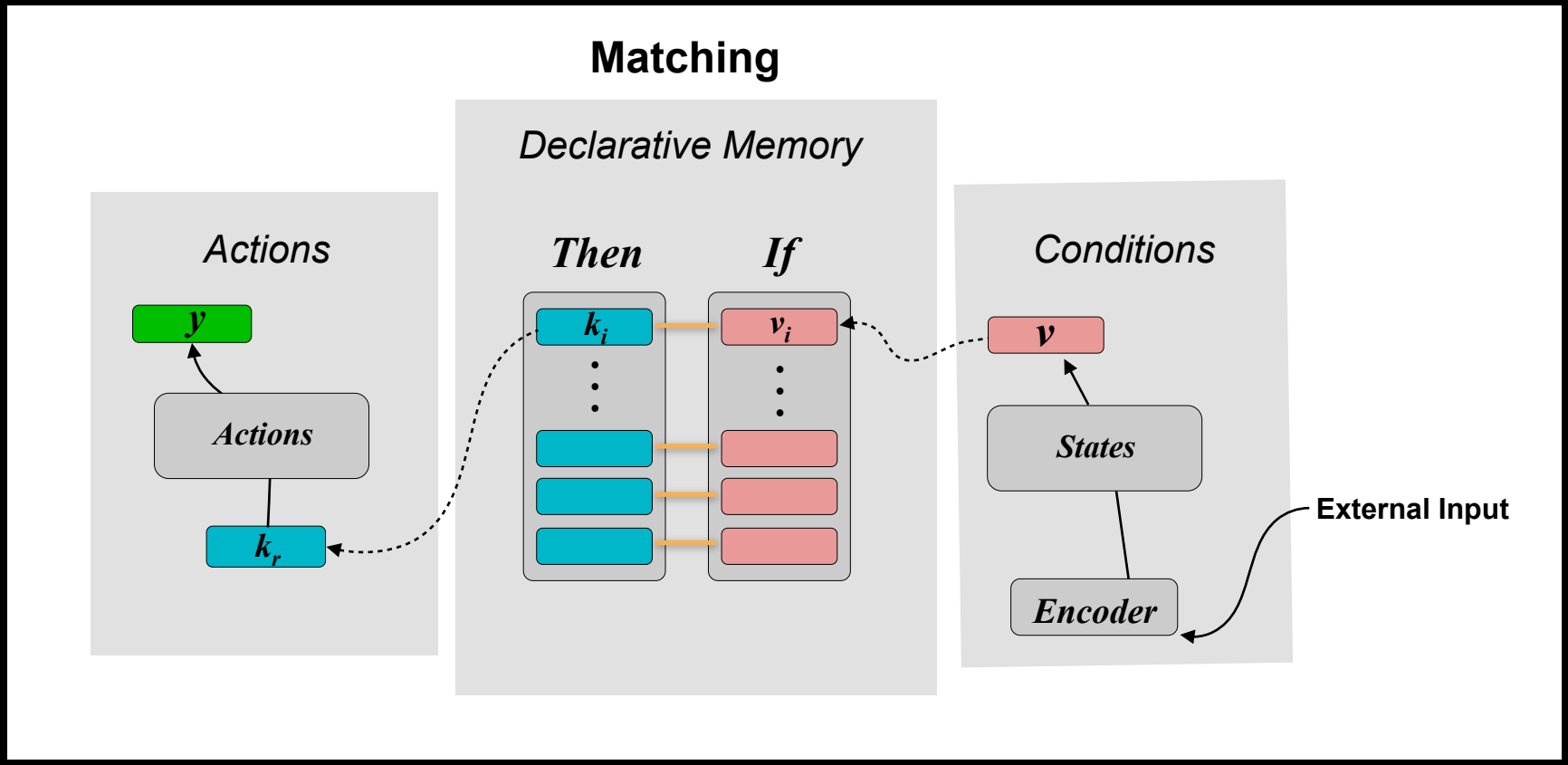
Production Systems with a Twist

Emergent Symbols Through Binding Network (ESBN)

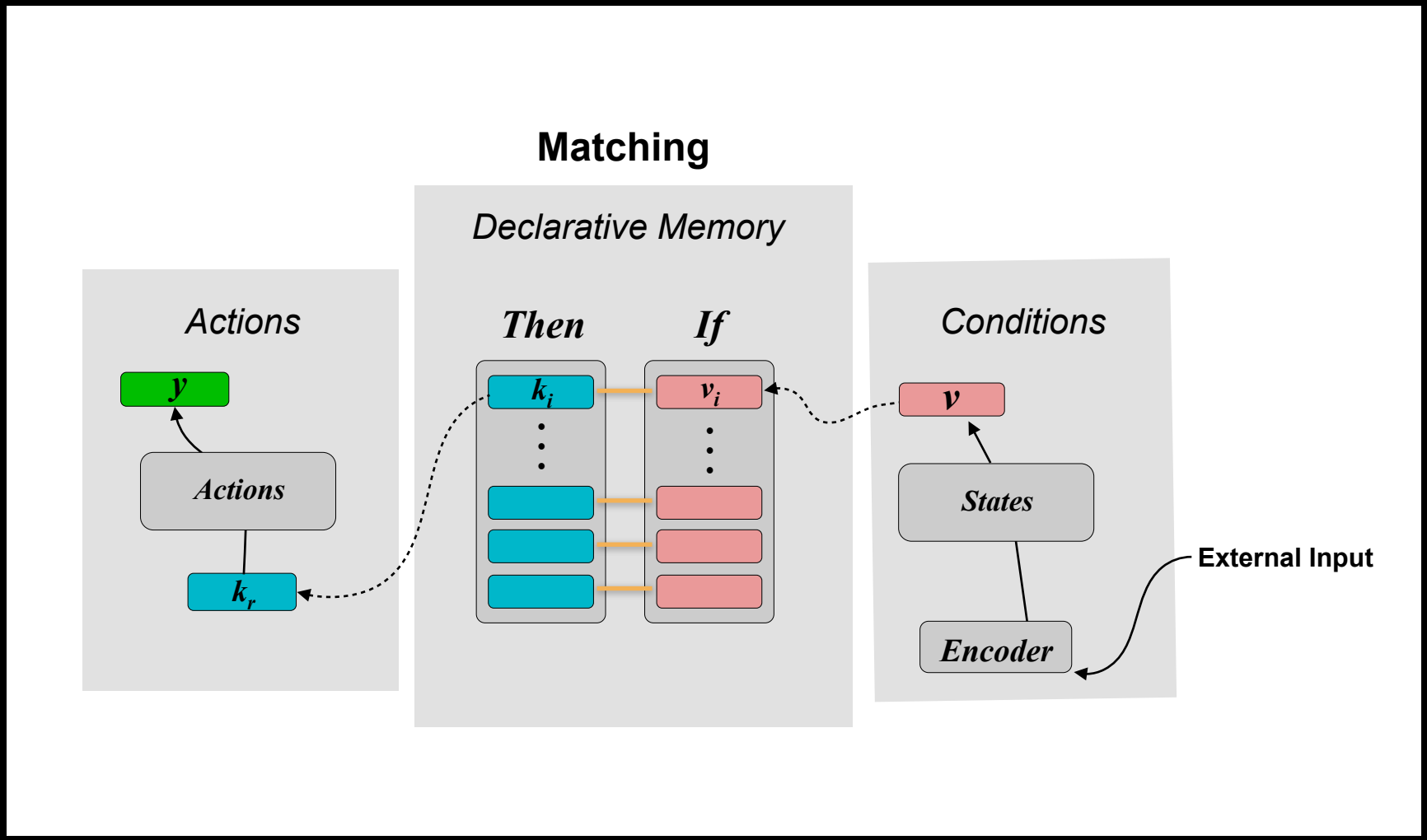
(Webb et al., ICLR 2021)



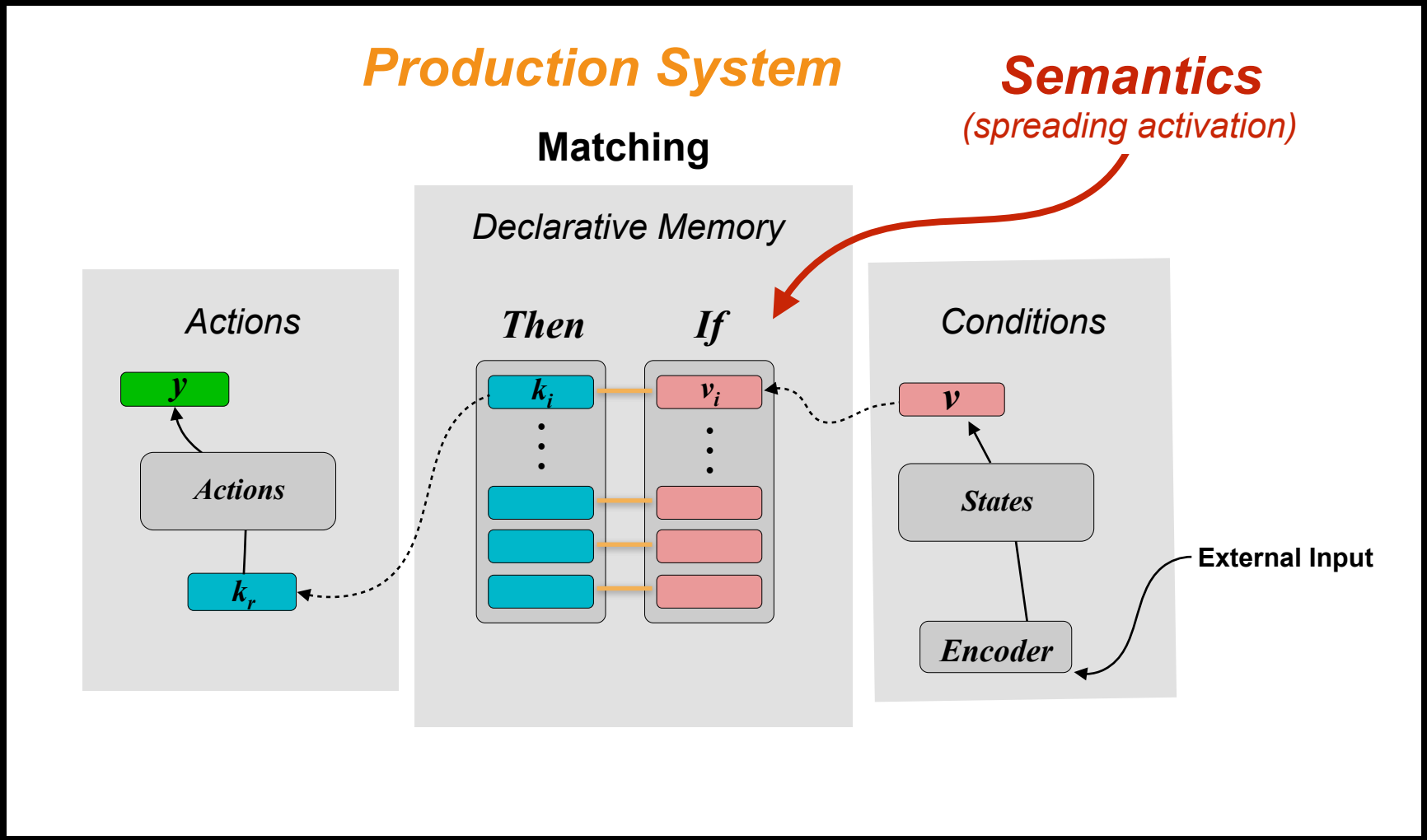
Production Systems with a Twist



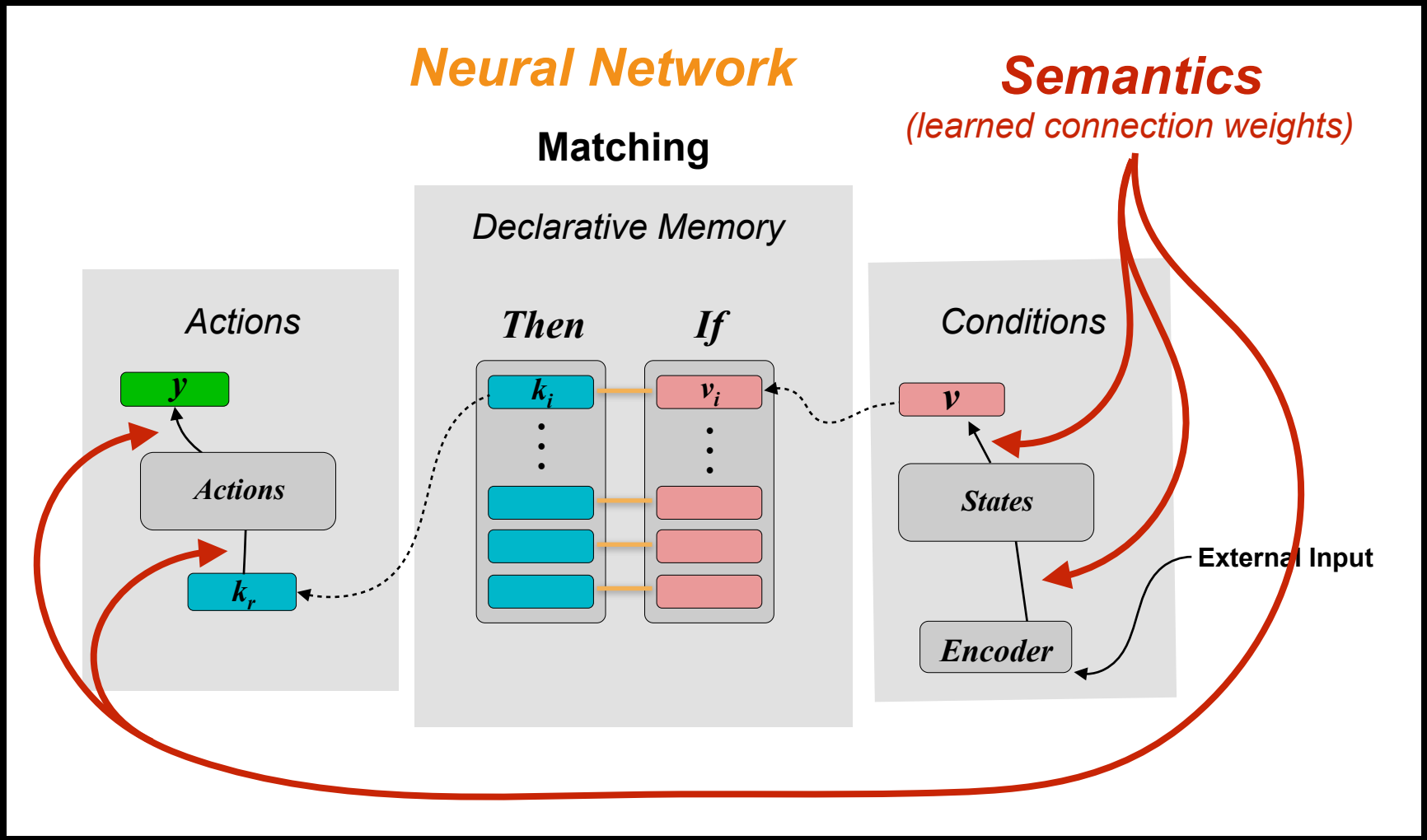
Production Systems with a Twist



Production Systems with a Twist



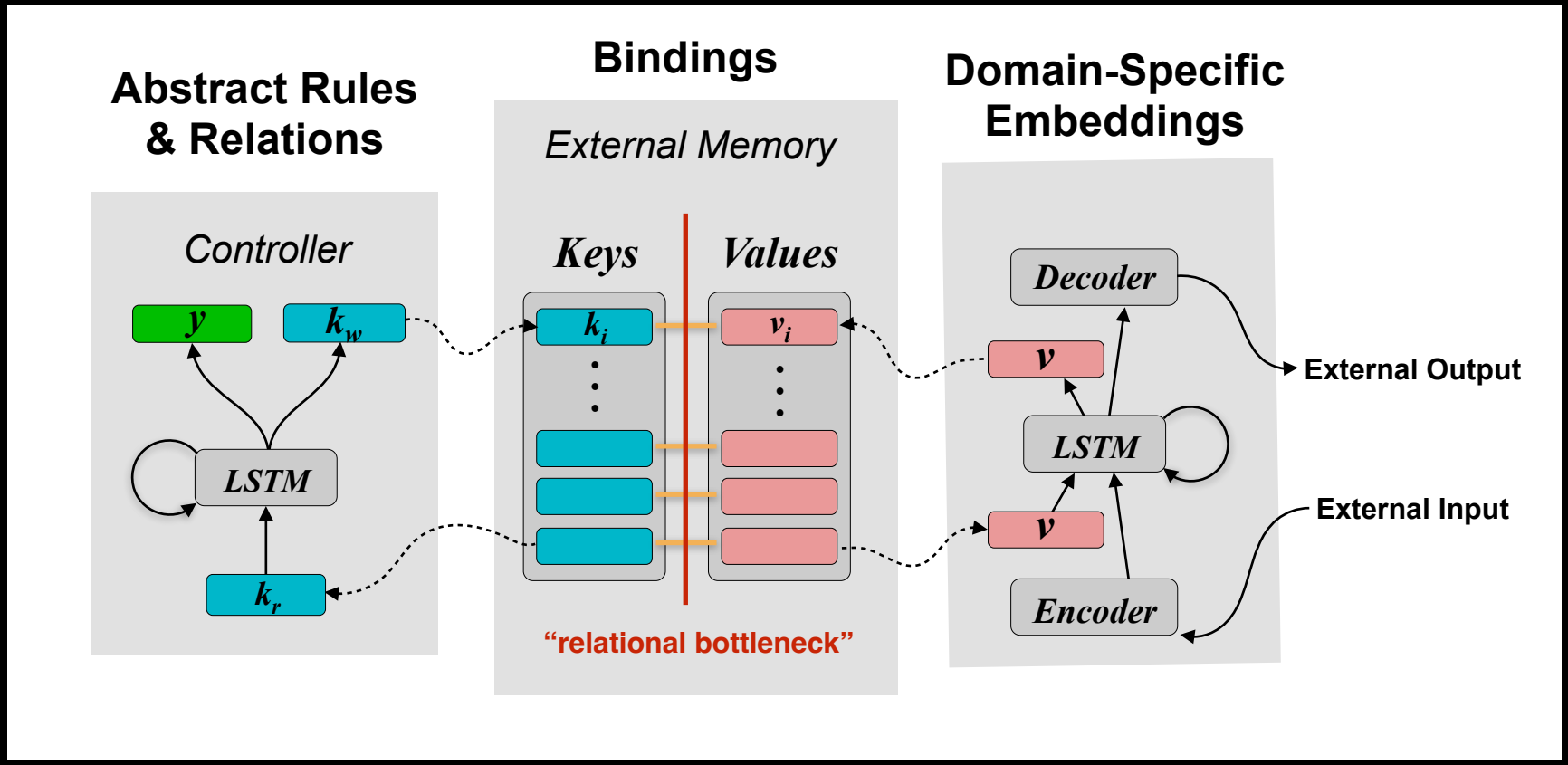
Production Systems with a Twist



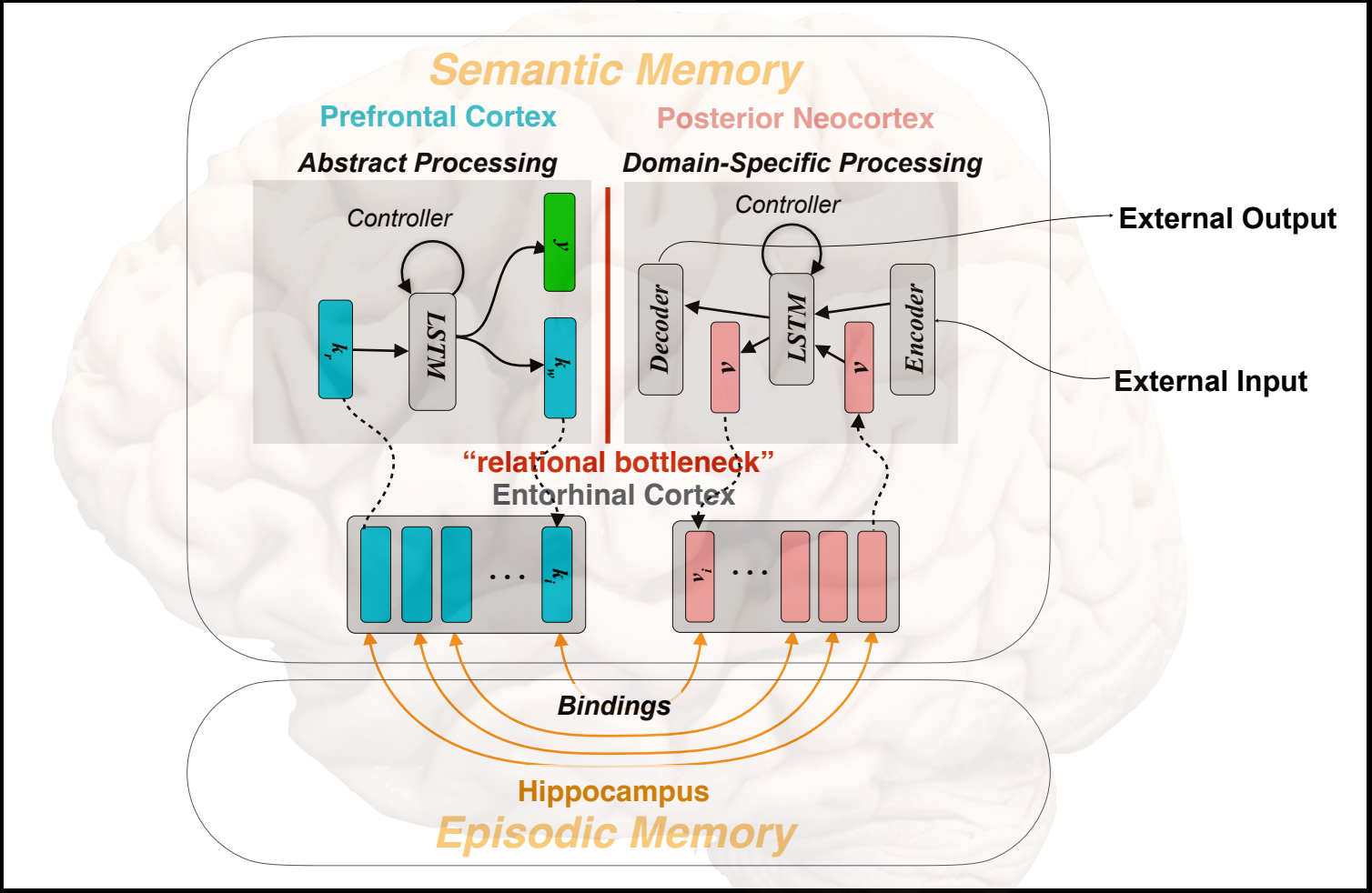
Comports with Architecture of the Brain

Emergent Symbols Through Binding Network (ESBN)

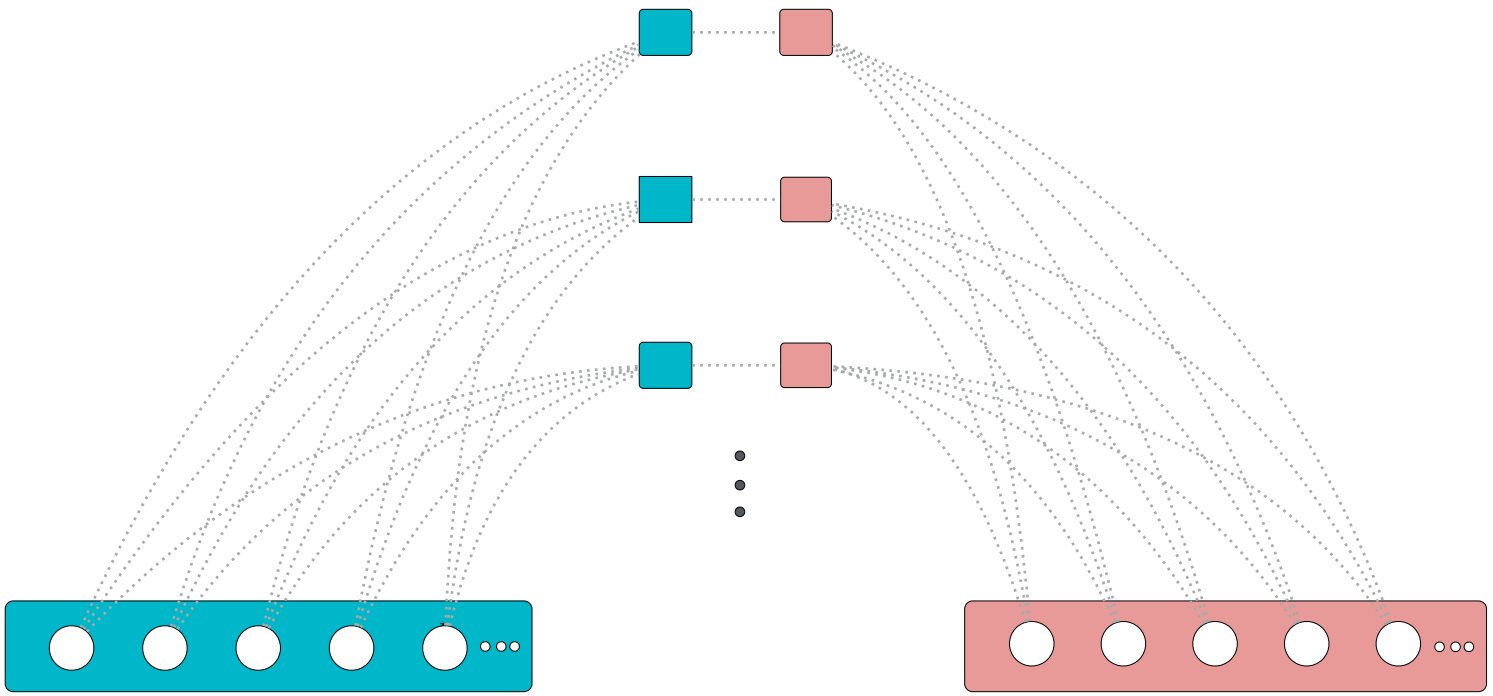
(Webb et al., ICLR 2021)



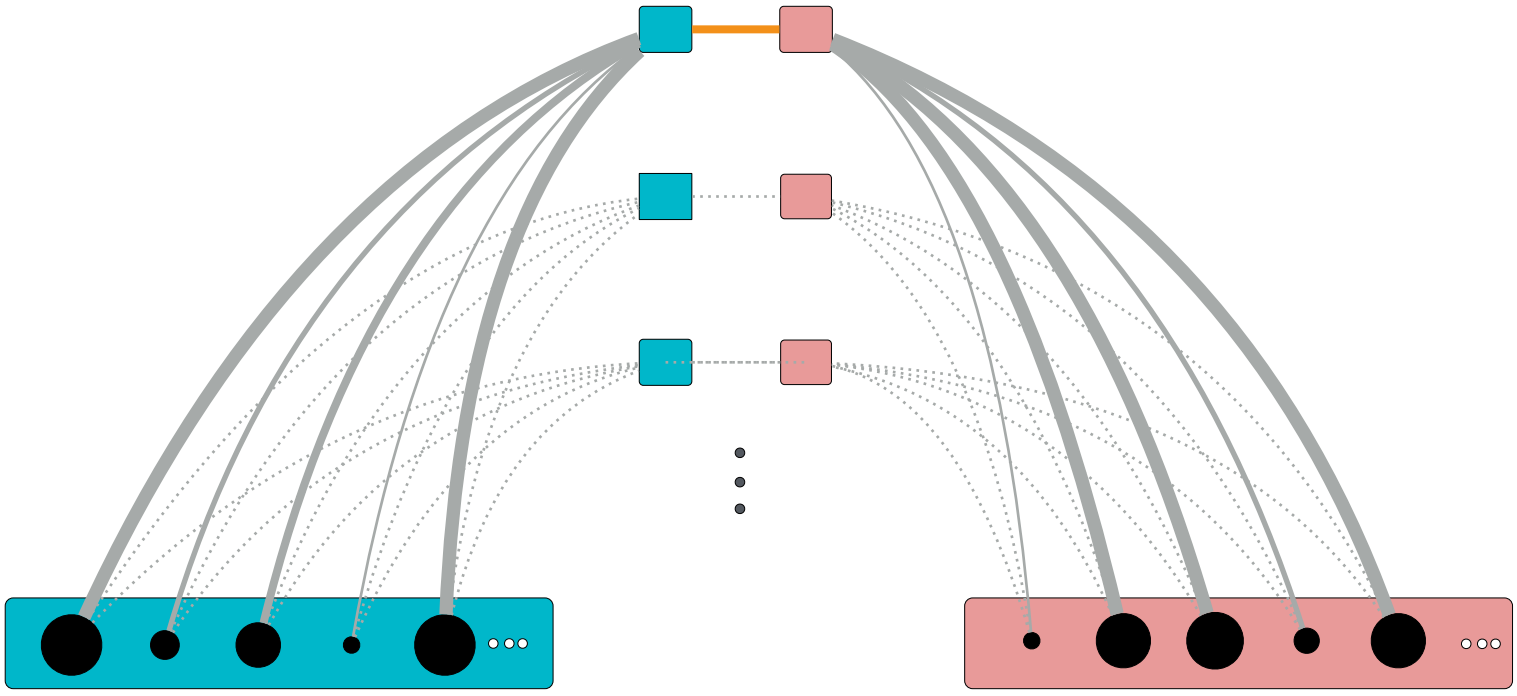
Relational Bottleneck in the Brain



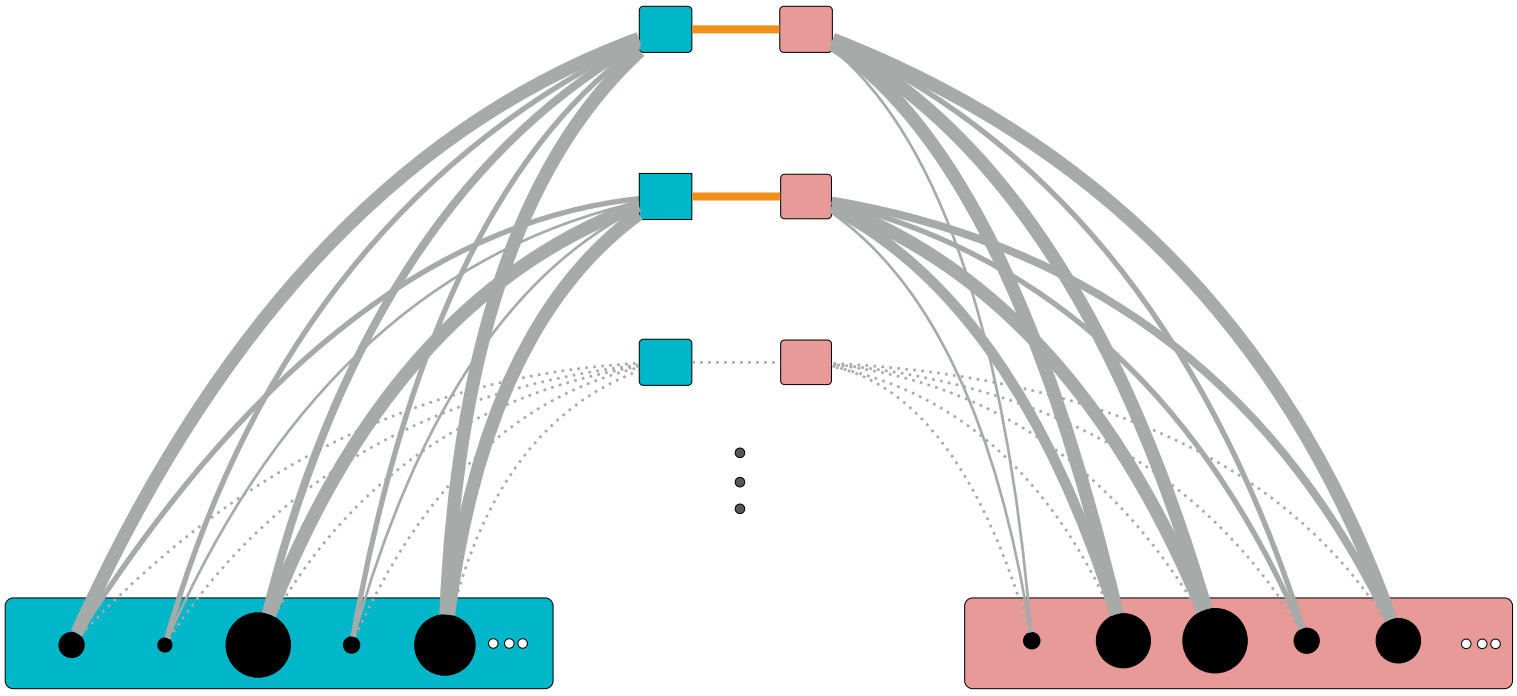
“Modern Hopfield Network”



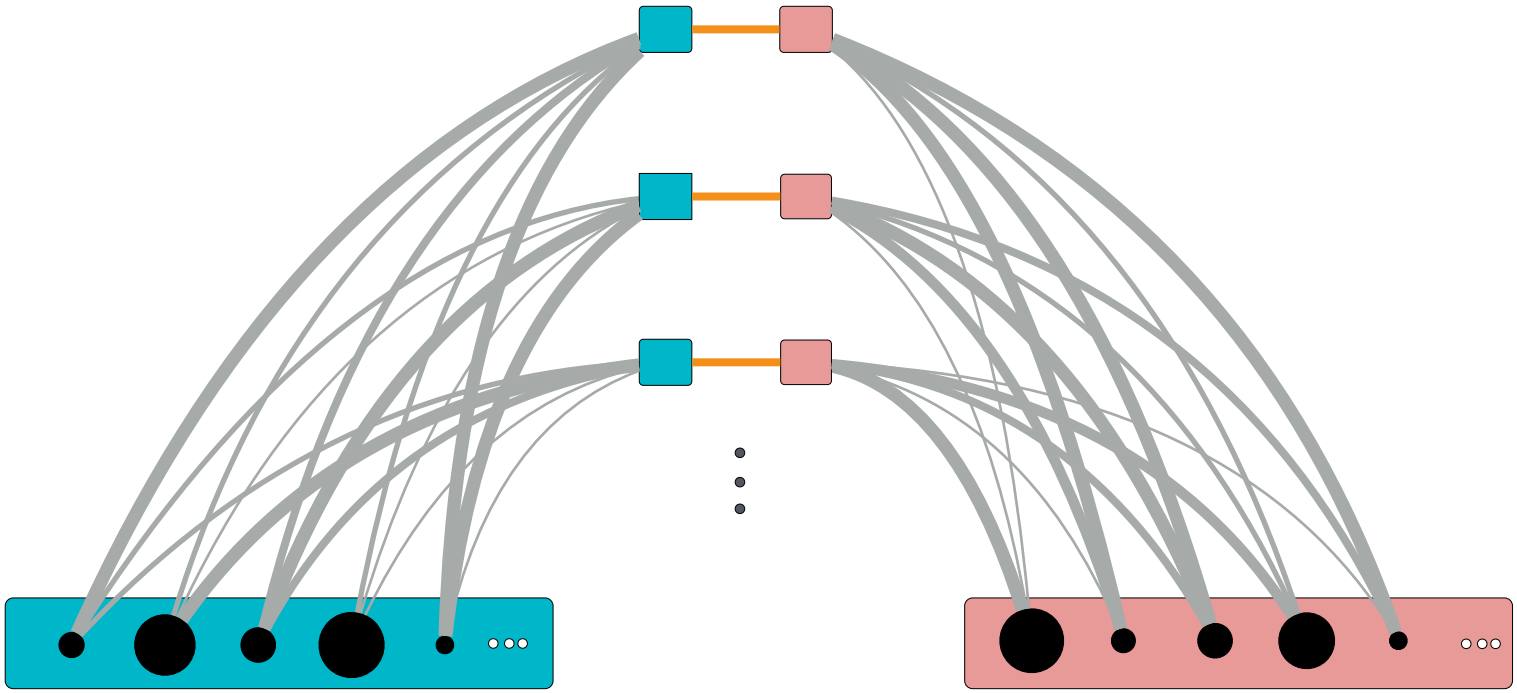
“Modern Hopfield Network”



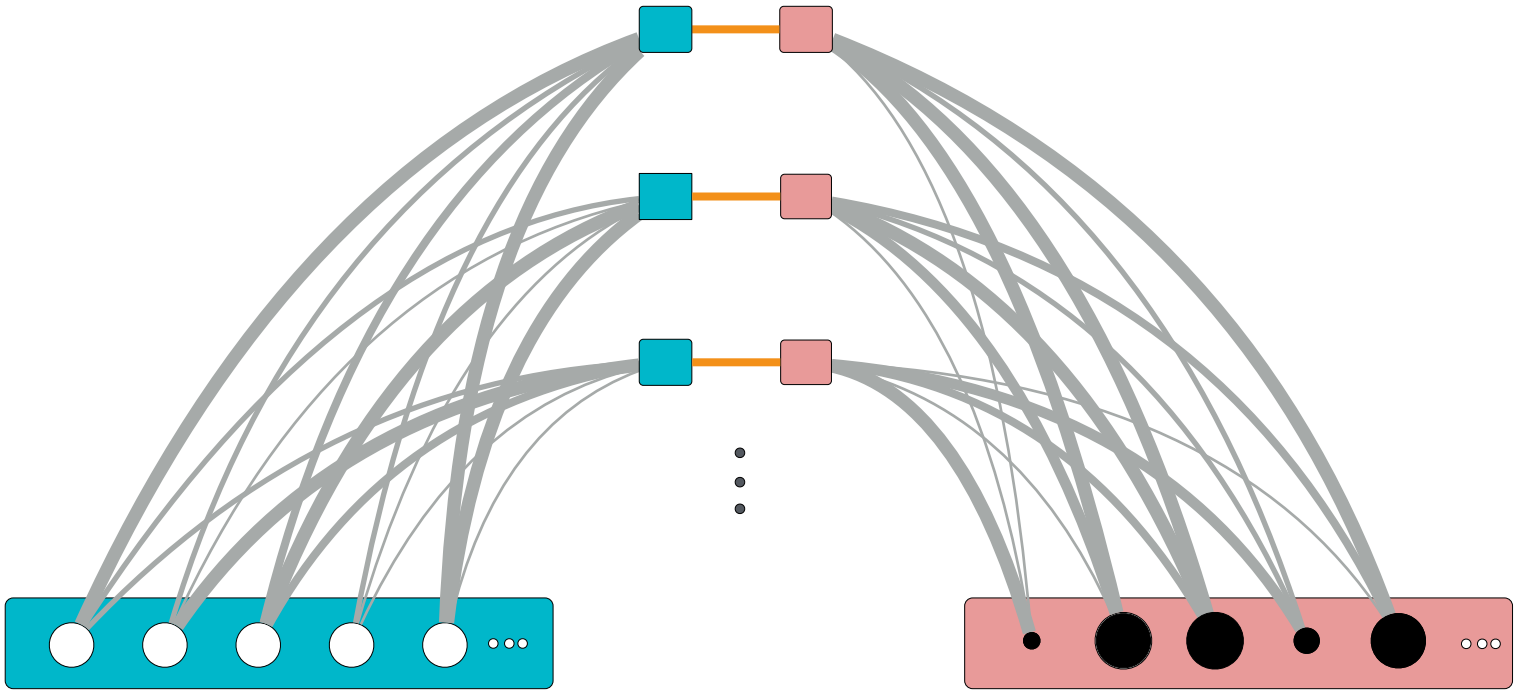
“Modern Hopfield Network”



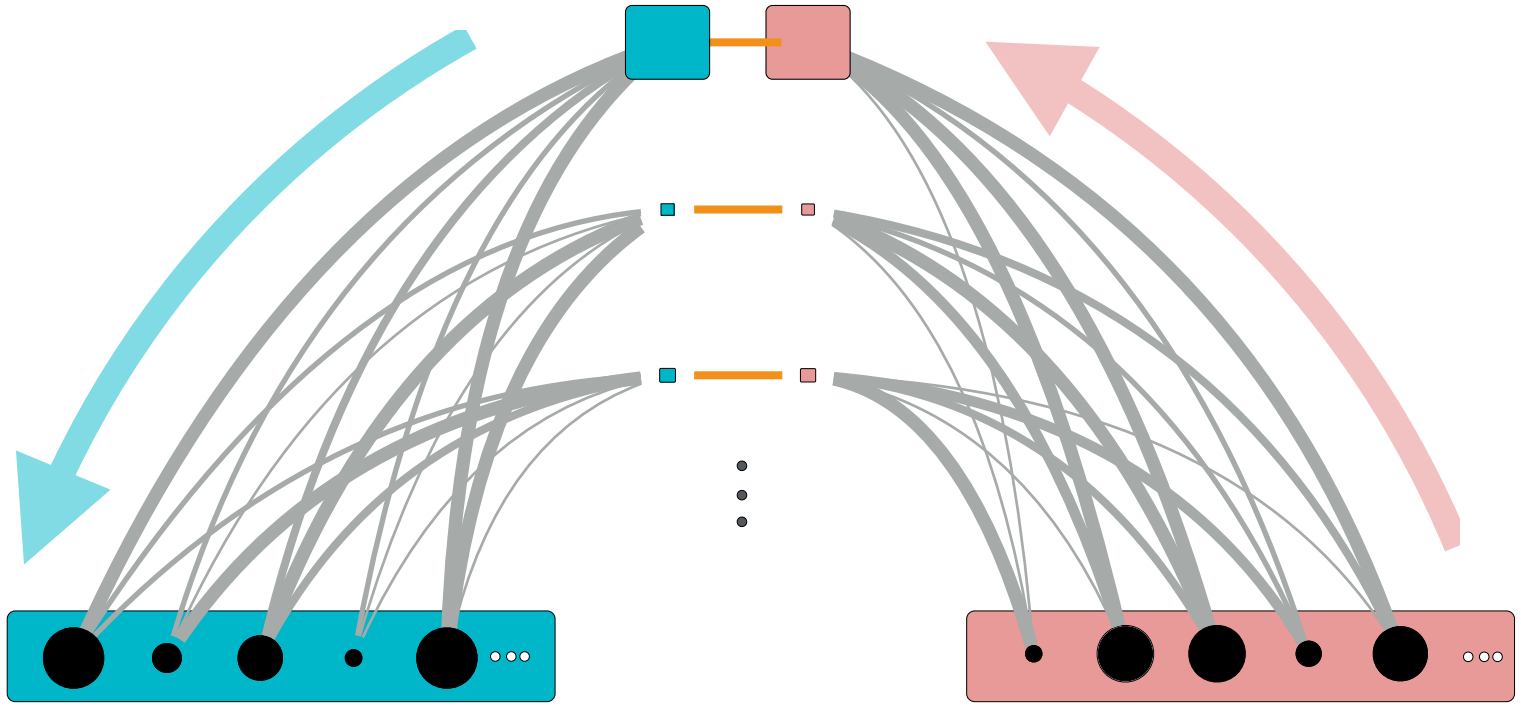
“Modern Hopfield Network”



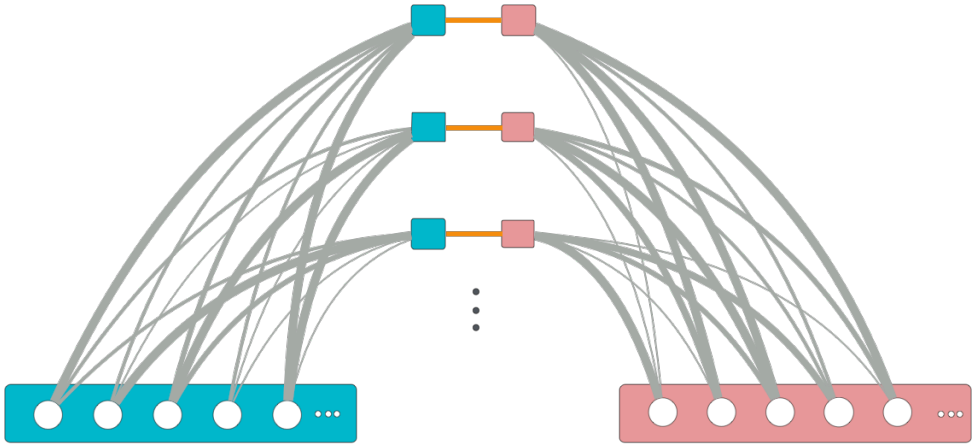
“Modern Hopfield Network”



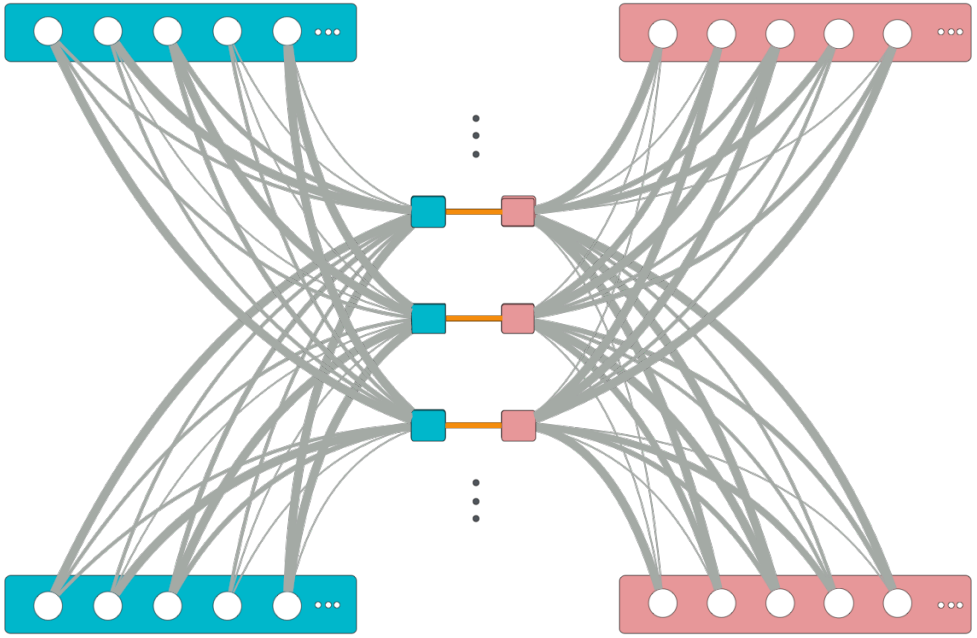
“Modern Hopfield Network”



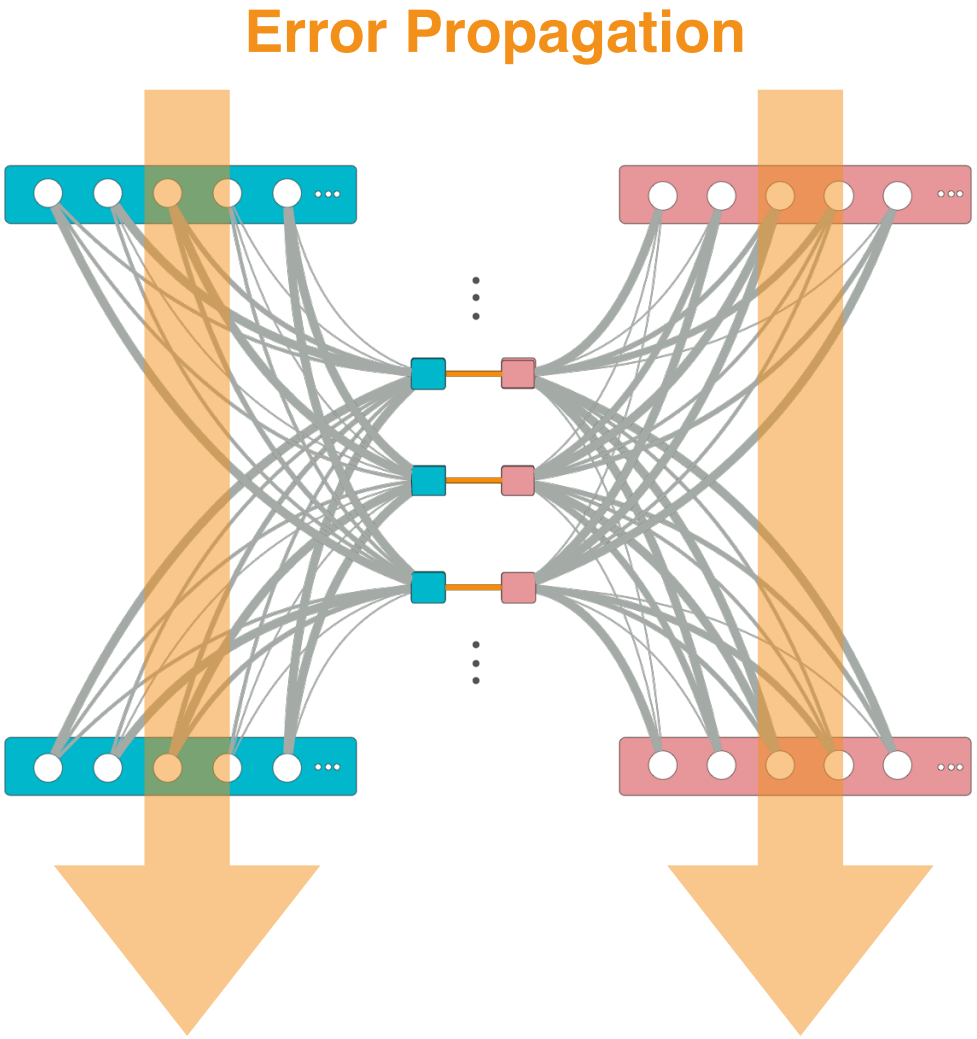
“Modern Hopfield Network”



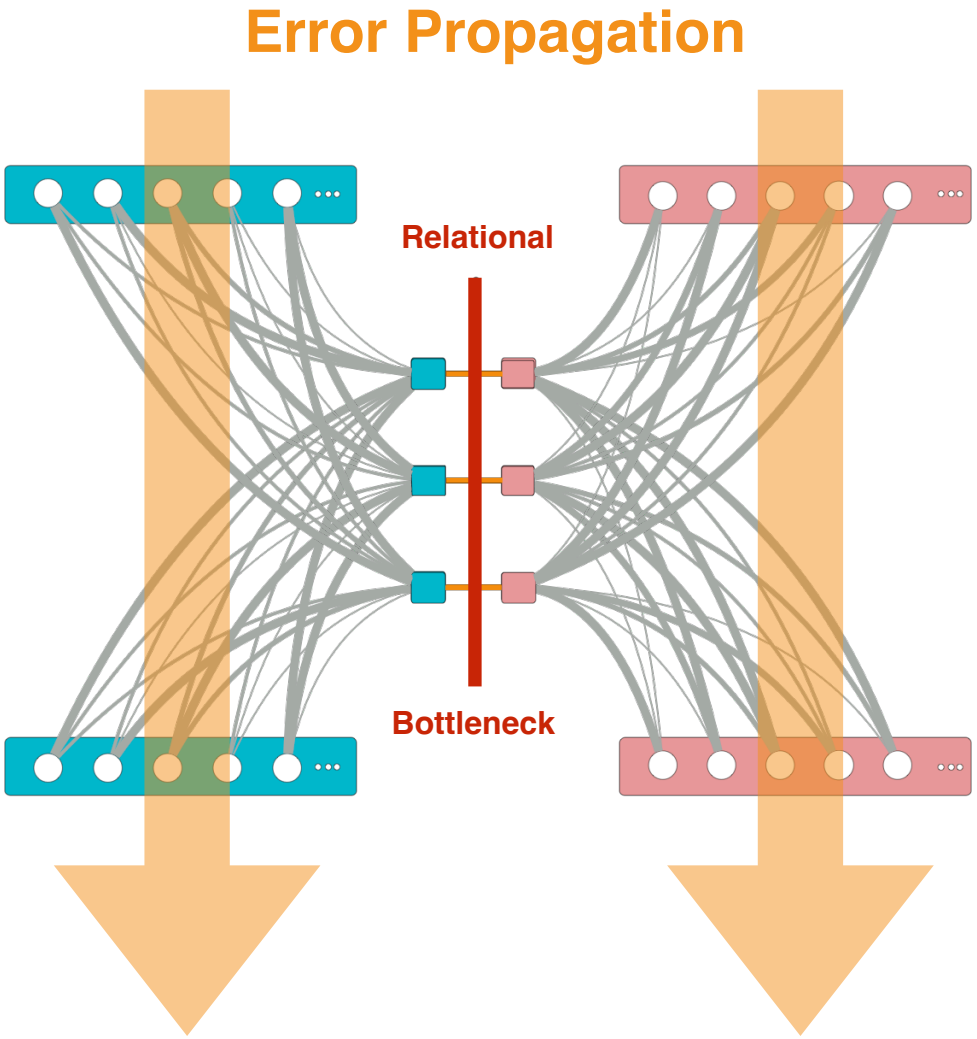
“Modern Hopfield Network”



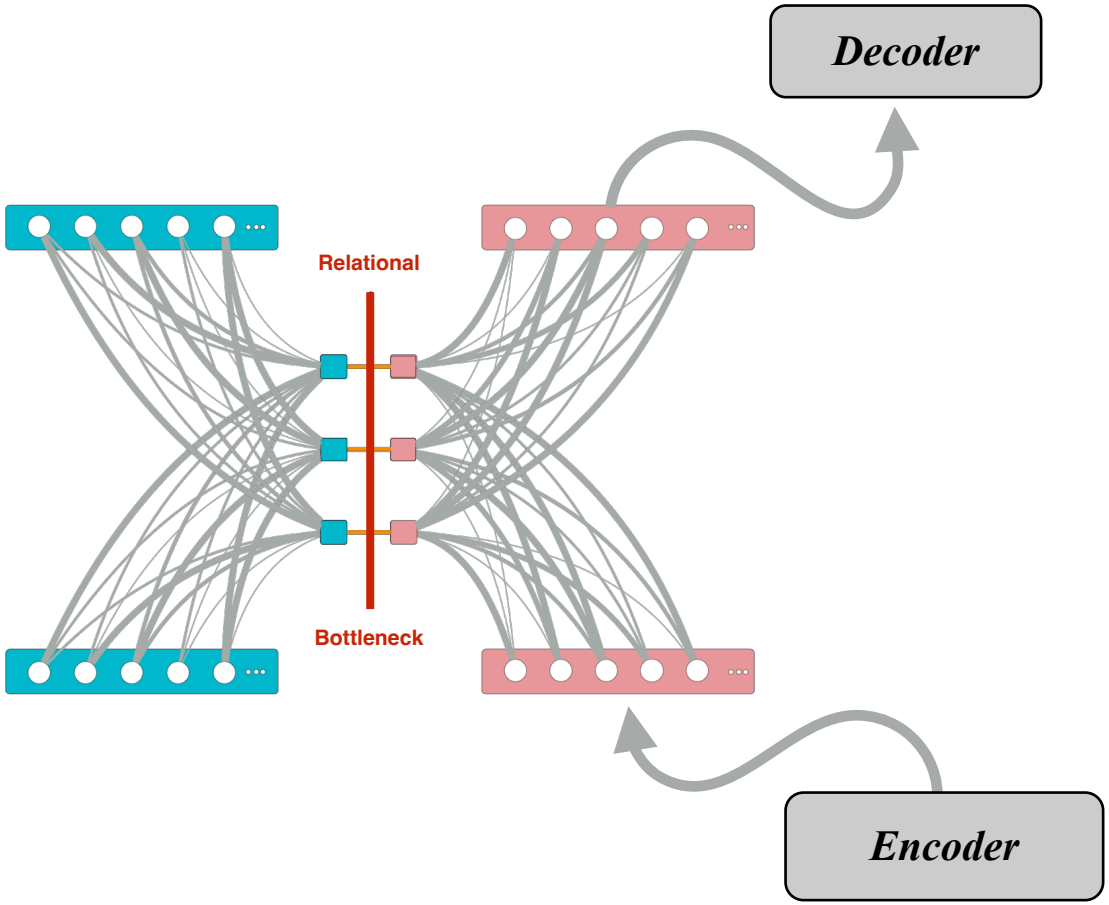
“Modern Hopfield Network”



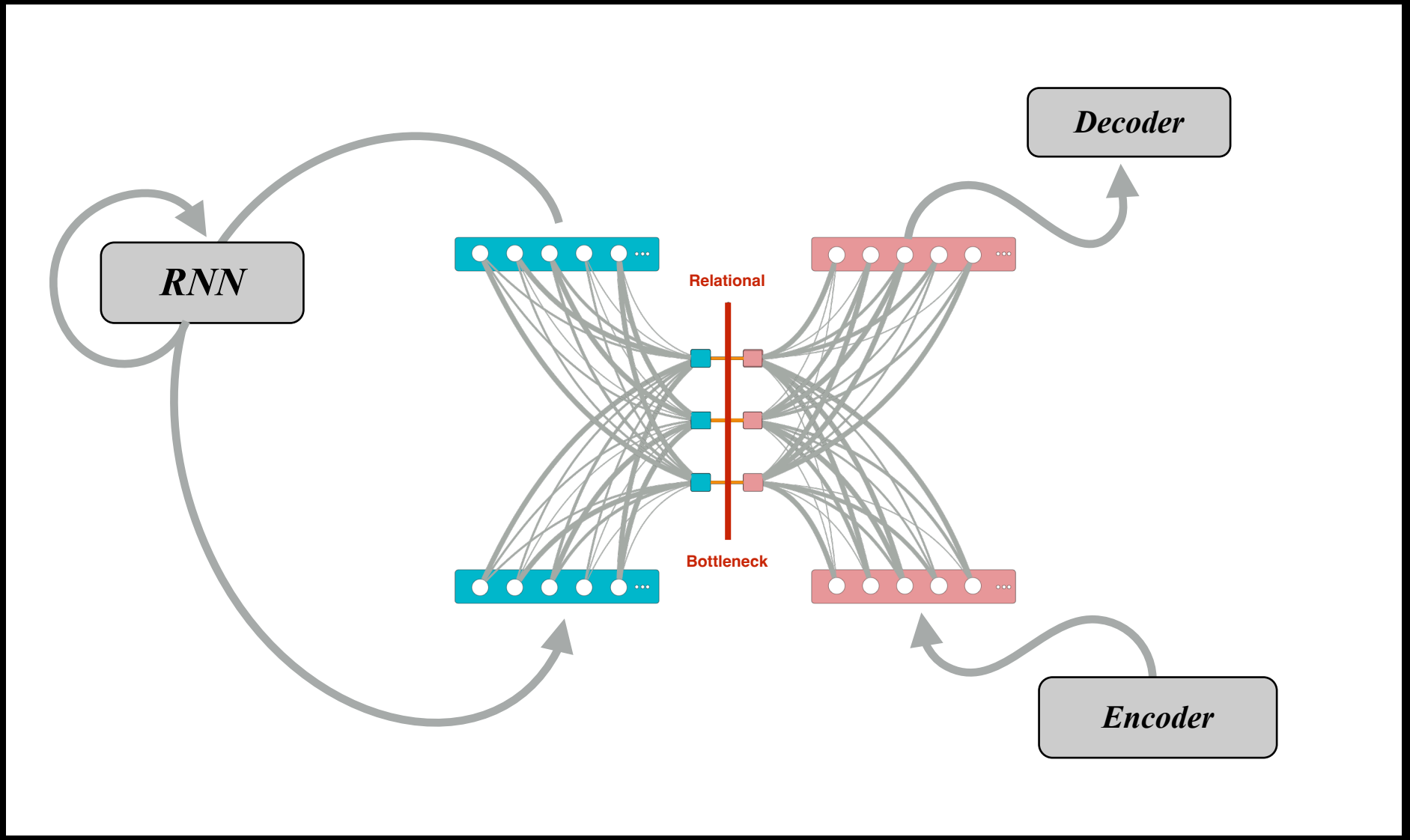
“Modern Hopfield Network”



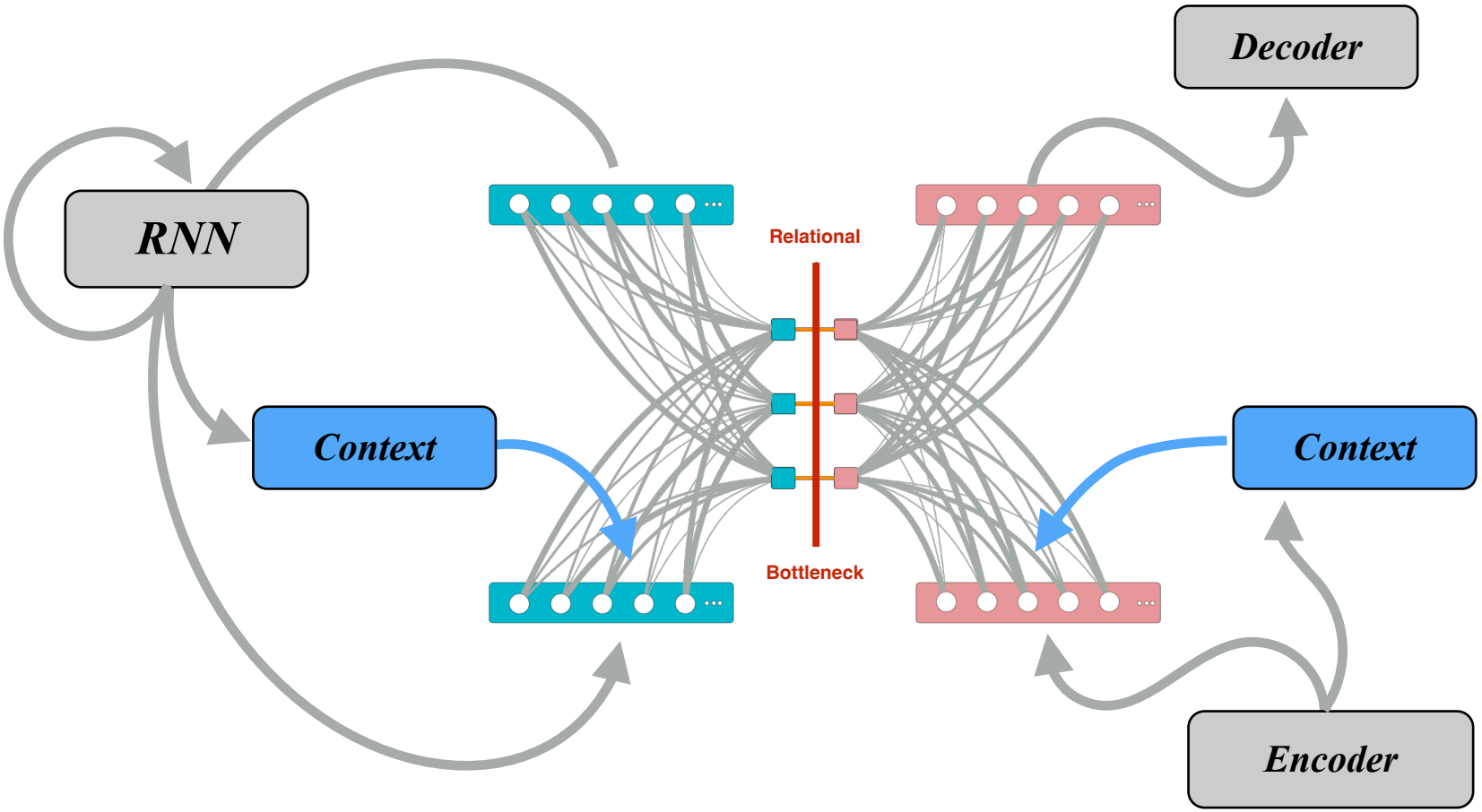
Full(er) Architecture



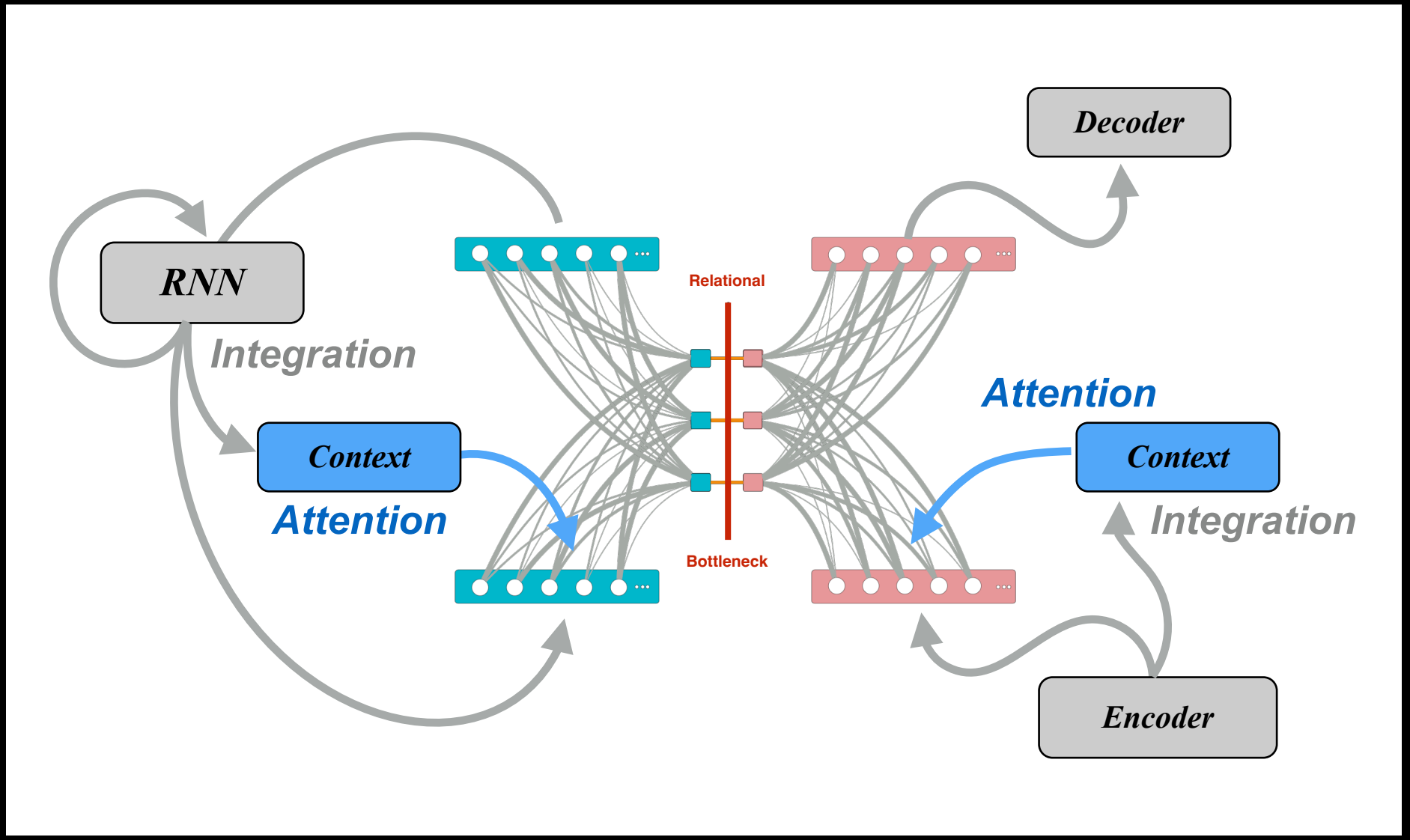
Full(er) Architecture



Full(er) Architecture



Full(er) Architecture



Full(er) Perspective

Full(er) Perspective

Abstraction requires inductive biases in the architecture

- **Integration**: ***formation*** **of context representations**

Full(er) Perspective

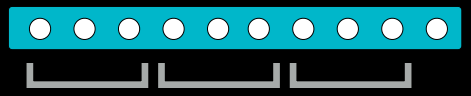
Abstraction requires inductive biases in the architecture

- **Integration**: *formation* of context representations
  - time averaging (lossy)

Full(er) Perspective

Abstraction requires inductive biases in the architecture

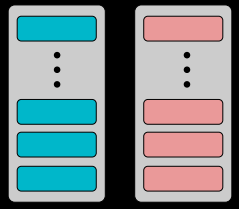
- **Integration**: *formation* of context representations
  - time averaging (lossy) 

  - windowed integration (discrete) 

Full(er) Perspective

Abstraction requires inductive biases in the architecture

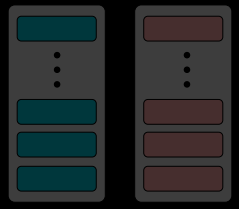
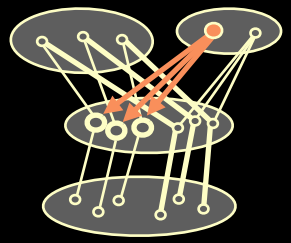
- **Integration**: *formation* of context representations
  - time averaging (lossy) 

  - windowed integration (discrete) 
  - episodic memory (lossless) 

Full(er) Perspective

Abstraction requires inductive biases in the architecture

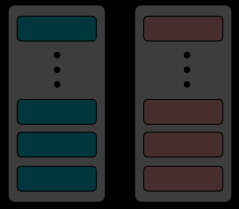
- Integration: *formation* of context representations
  - time averaging (lossy) 

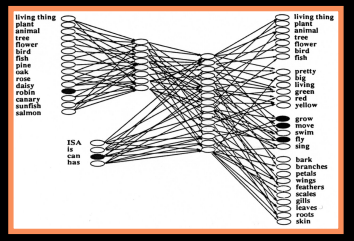
  - windowed integration (discrete) 
  - episodic memory (lossless) 
- Attention: *use* of context representations
  - selection of mappings 

Full(er) Perspective

Abstraction requires inductive biases in the architecture

- Integration: *formation* of context representations
  - time averaging (lossy) 

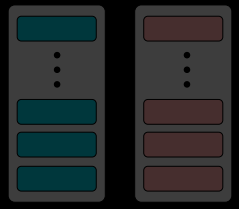
  - windowed integration (discrete) 
  - episodic memory (lossless) 
- **Attention: *use* of context representations**
  - selection of mappings
  - **shaping of statistical structure**

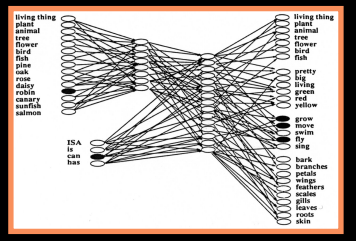


Full(er) Perspective

Abstraction requires inductive biases in the architecture

- Integration: *formation* of context representations
  - time averaging (lossy) 

  - windowed integration (discrete) 
  - episodic memory (lossless) 
- **Attention: *use* of context representations**
  - selection of mappings
  - **shaping of statistical structure**



Integration, Attention & Semantics (ISC-CI)

Human Semantic Anomalies / Idiosyncrasies

- **Similarity Judgments:**
  - **Order effects: asymetric similarity judgements**
    How similar is a donkey to a horse? *Similar*
    How similar is a horse to a donkey? *Less so*
  - **Multialternative effects: reversal of simiarlity judgments**
    Which is Jamaica most similar to: England, Poland or Cuba *Cuba*
    Which is Jamaica most similar to: England, Russia or Cuba *England*
  - **Triangle inequality effects: violations of transitive inference**
    [Nurse : Patient] :: [Mother : Child]
    [Mother : Child] :: [Frog : Tadpole]
    [Nurse : Patient] :: [Frog : Tadpole]

Human Semantic Anomalies / Idiosyncrasies

- **Similarity Judgments**
- **Category Judgements** *(Inductive Inference)*:

  Given each of the following sets, which is more likely to be a member of the category:

  - **Premise-conclusion similarity:**
    {crows} → ravens or *{crows} → robins*
  - **Conclusion typicality:**
    {crows, ravens} → robins or *{crows, ravens} → penguins*
  - **Premise diversity:**
    {crows, robins} → sparrows or *{crows, ravens} → sparrows*
  - **In-category monotonicity** *(support spans query)*
    {crows, robins} → sparrows or *{crows} → sparrows*
  - **In-category non-monotonicity** *(support is narrower than query)*
    {brown bears, grizzly bears} → buffalo or *{brown bears} → buffalo*
  - **Cross-category non-monotonicity** *(support is broader than query)*
    {flies, orangutans} → bees or *{flies} → bees*

Human Semantic Anomalies / Idiosyncrasies

Human Semantic Anomalies / Idiosyncrasies

- **Similarity Judgments**
- **Category Judgements** *(Inductive Inference)*
- **Theories:**
  - **Metric Theories:** ***parametric distances*** of representations in a high dimension vector space *(Smith, Shobin, Rips, 1974; Miklov, 2013 - e.g. ,Word2Vec, ~ISC)*

Human Semantic Anomalies / Idiosyncrasies

- **Similarity Judgments**
- **Category Judgements** *(Inductive Inference)*
- **Theories:**
  - **Metric Theories:** ***parametric distances*** of representations in a high dimension vector space *(Smith, Shobin, Rips, 1974; Miklov, 2013 - e.g. ,Word2Vec, ~ISC)*
  - **Feature Contrast / Coverage Theories:** ***discrete (set theoretic) intersection*** / overlap *(Tversky, 1977; Osherson et al., 1990)*

Human Semantic Anomalies / Idiosyncrasies

- **Similarity Judgments**
- **Category Judgements** *(Inductive Inference)*
- **Theories:**

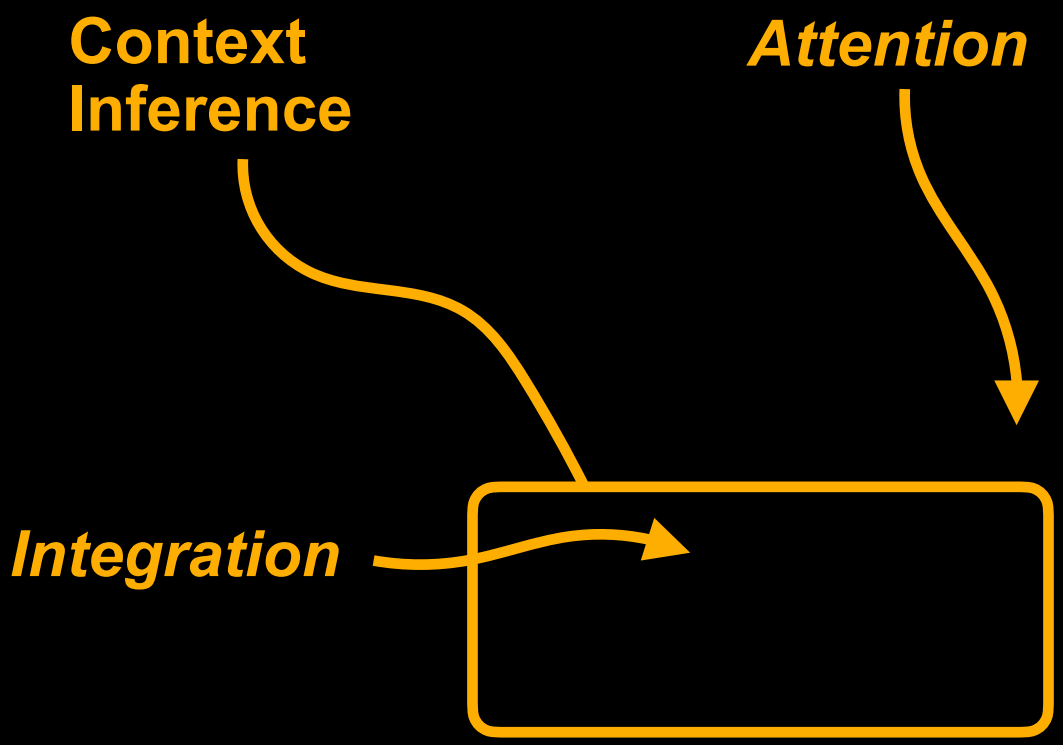
  - **Metric Theories:** ***parametric distances*** of representations in a high dimension vector space
    (Smith, Shobin, Rips, 1974; Miklov, 2013 - e.g. ,Word2Vec, ~ISC)

  - **Feature Contrast / Coverage Theories:** ***discrete (set theoretic) intersection*** / overlap
    (Tversky, 1977; Osherson et al., 1990)

  - **Bayesian Inference Theories**: likelihood relative to prior
    (Xu & Tenenbaum, 2007; Griffiths et al., 2010)

Integrated Semantics and Control — Context *Inference* (ISC-CI)

Giallanza, Campbell, Rogers & Cohen (under review)



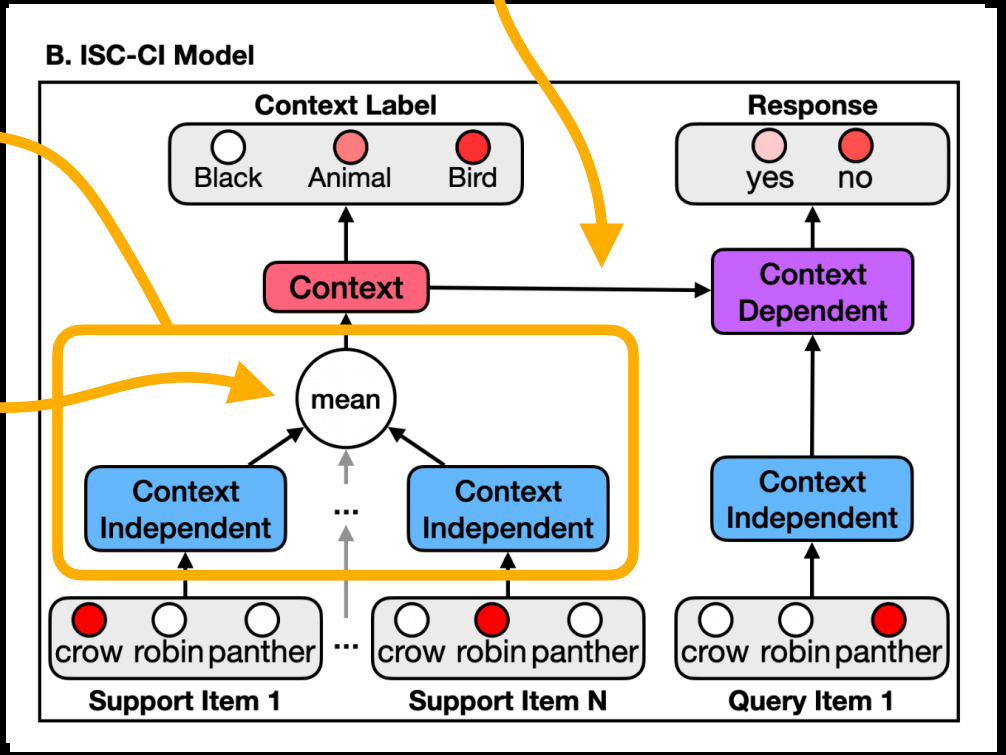
Integrated Semantics and Control — Context *Inference* (ISC-CI)

Giallanza, Campbell, Rogers & Cohen (under review)

Context Inference

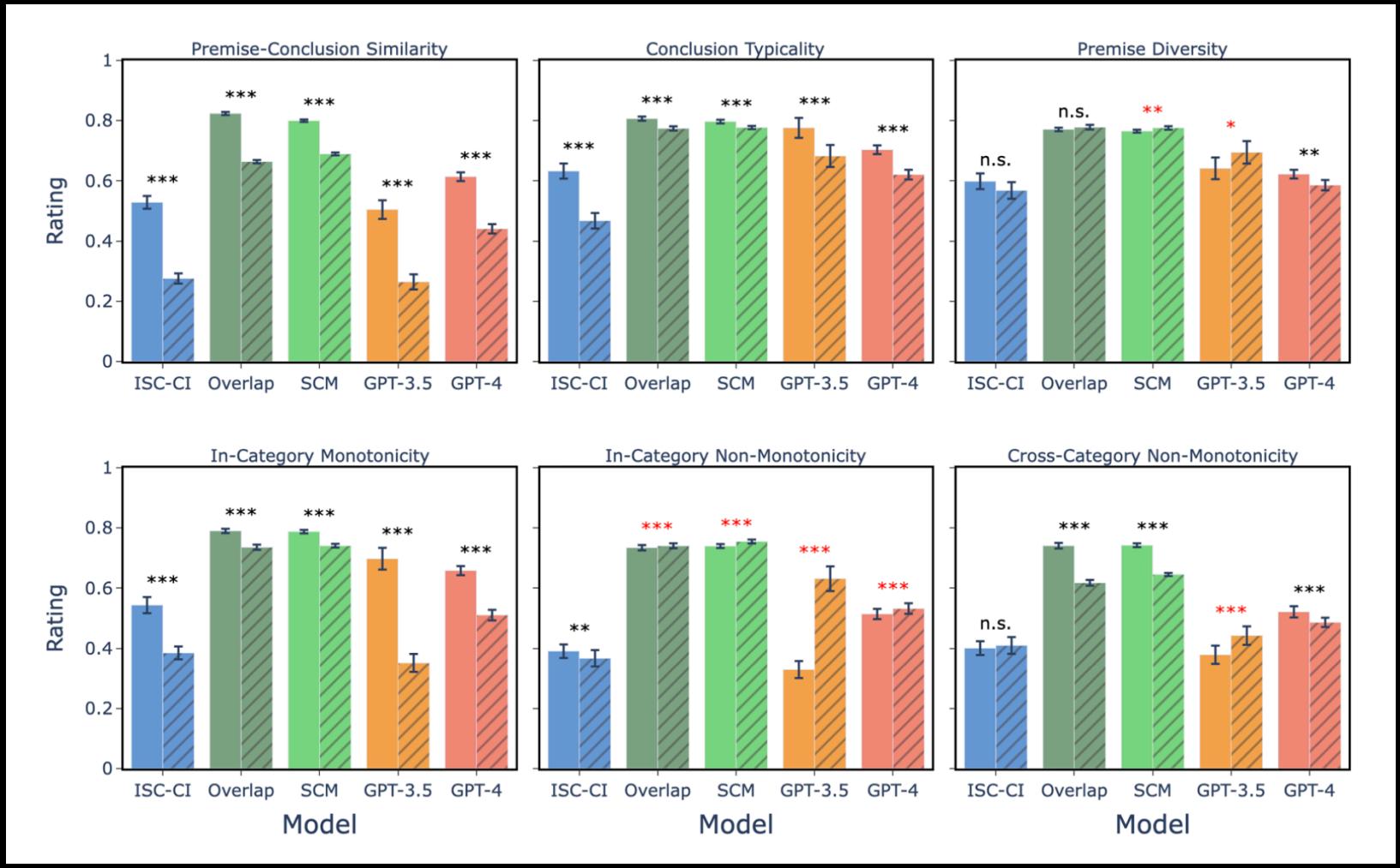
Attention

Integration



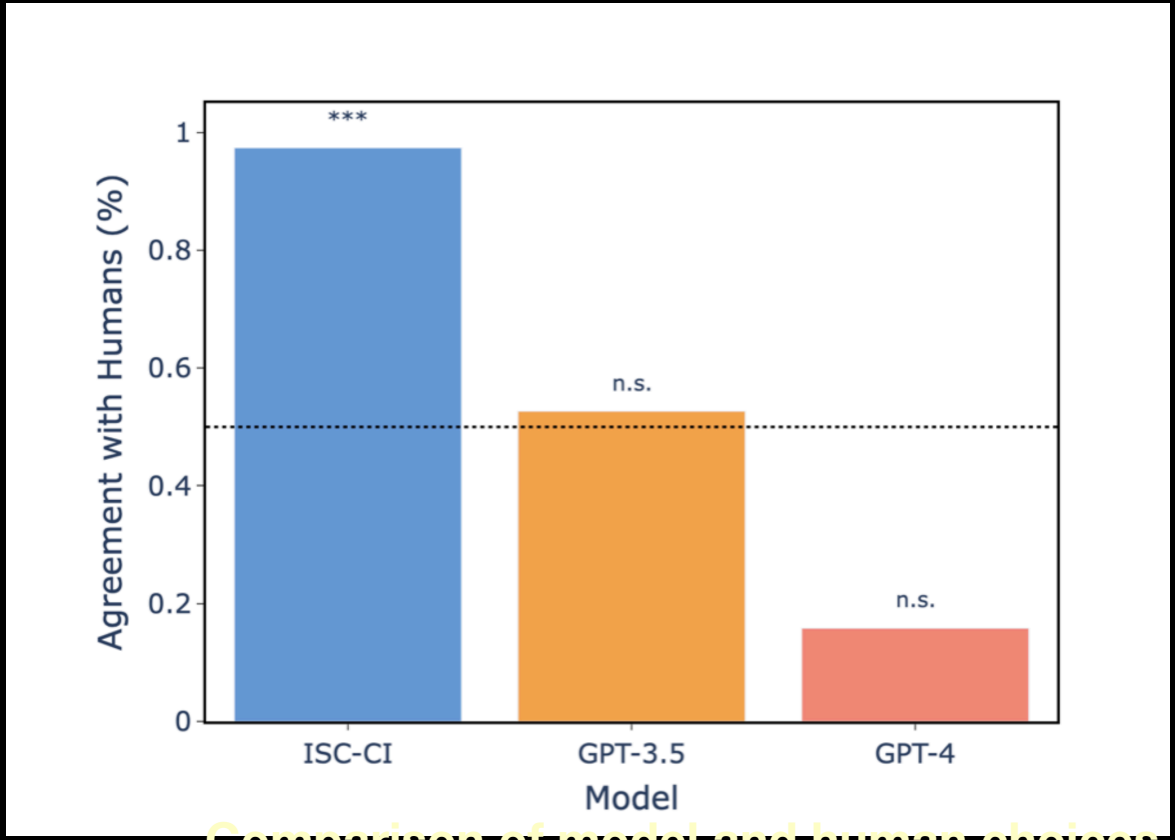
Integrated Semantics and Control — Context *Inference* (ISC-CI)

Giallanza, Campbell, Rogers & Cohen (under review)



Integrated Semantics and Control — Context *Inference* (ISC-CI)

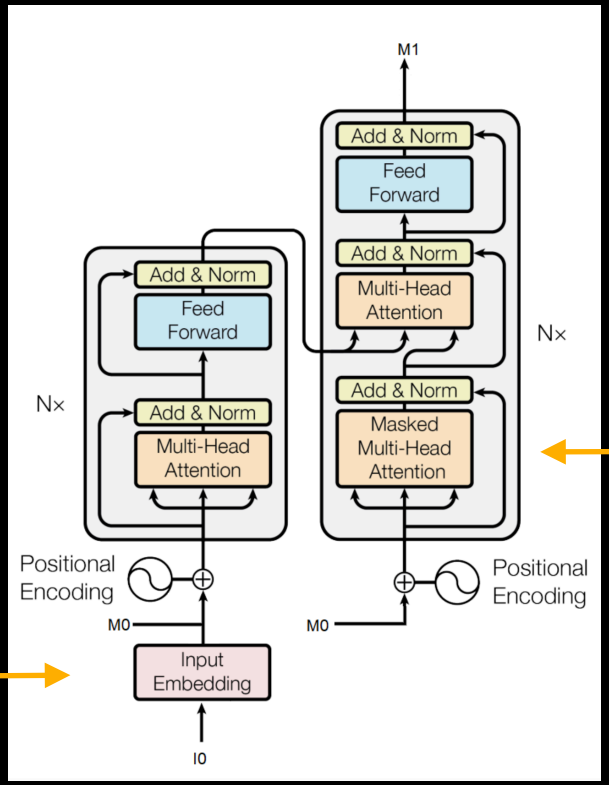
Giallanza, Campbell, Rogers & Cohen (under review)



Comparison of model and human choices in multi-alternative similarity judgments *(Tversky & Gati, 1978)*

Transformer Architecture

Output

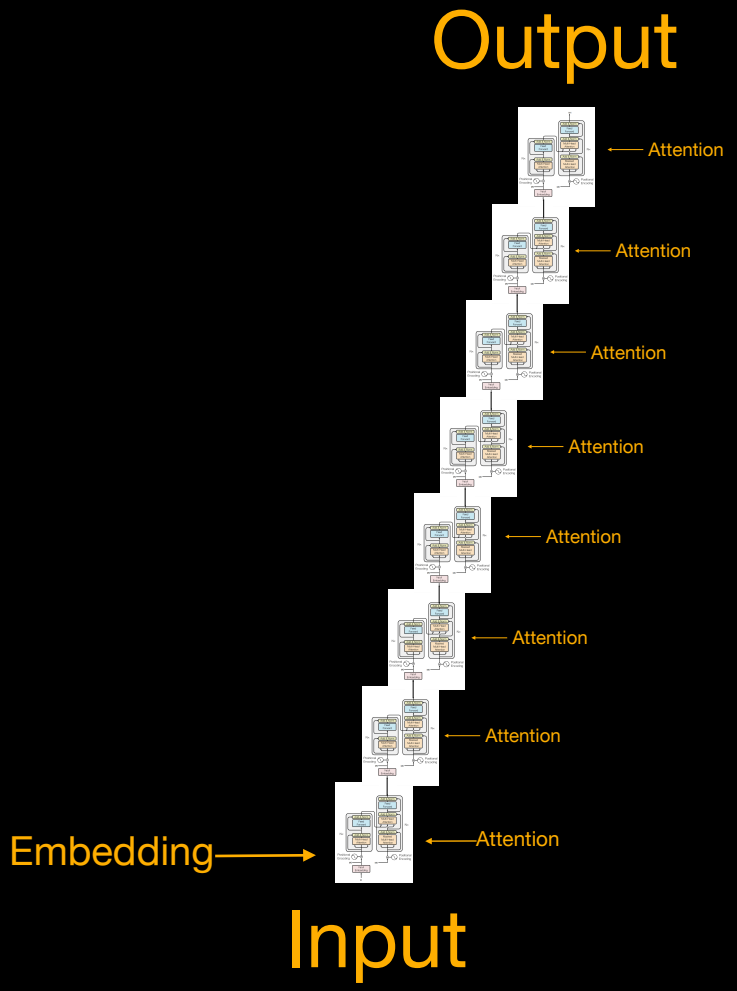


Attention

Embedding

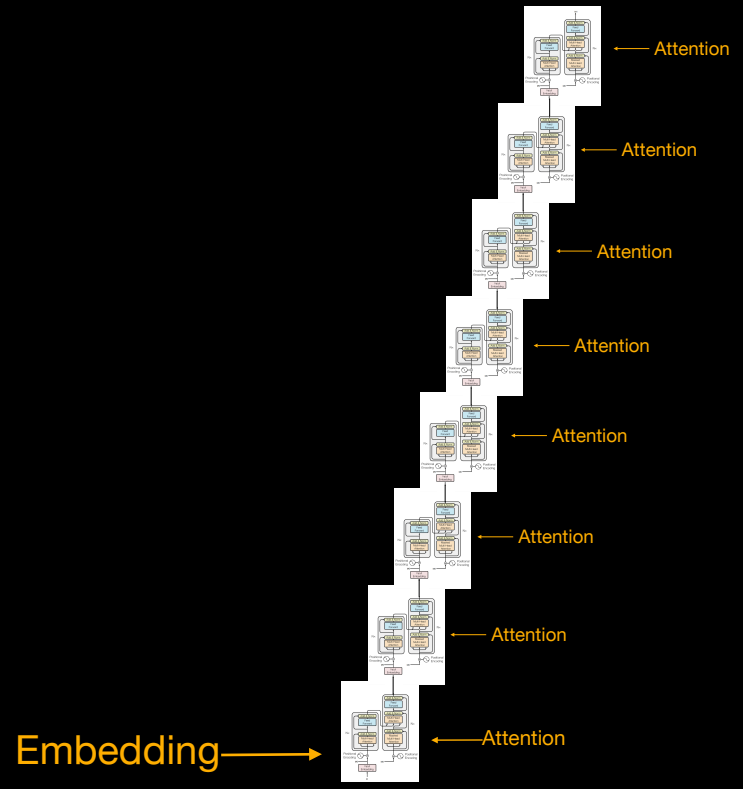
Input

Large Langauge / Vision Models



Large Langauge / Vision Models

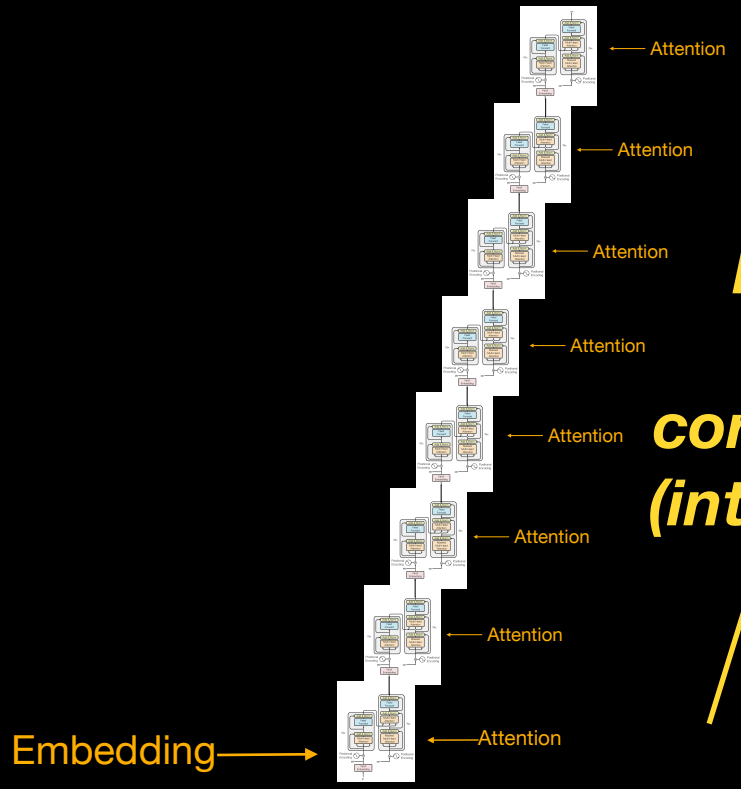
Output



Input *5,000 years worth of data!*

Large Langauge / Vision Models

Output

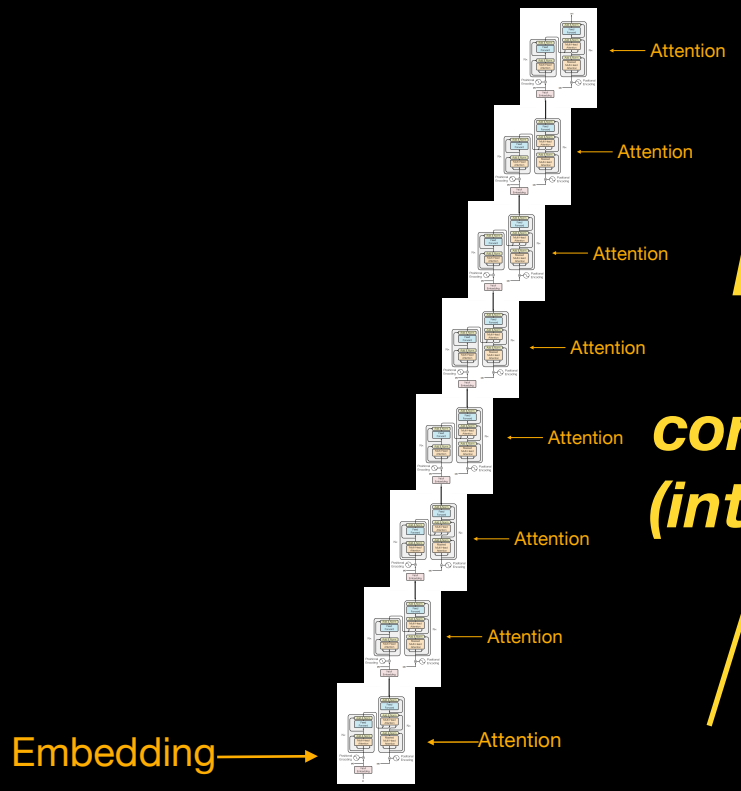


Higher order correlations (integration)

Input ***5,000 years worth of data!***

Large Langauge / Vision Models

Output



Higher order correlations (integration)

⇒ Abstraction

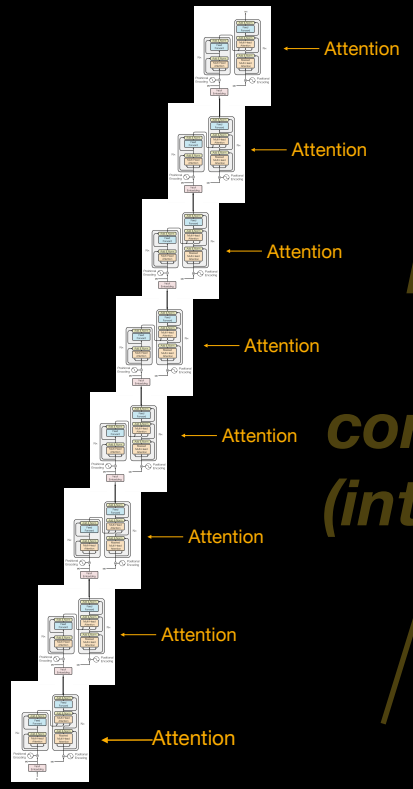
Input

5,000 years worth of data!

Large Langauge / Vision Models

Statistics on Steroids!

Output



Input

Higher order correlations (integration)

⇒ Abstraction

Acknowledgements

Steven Frankland
Dartmouth University

Tyler Gallant
Princeton Psychology

Taylor Webb
Microsoft Research

Shanka Mondal
Princeton Electrical Engineering

John Lafferty
Yale Computer Science

Simon Segert
Princeton Neuroscience

Kamesh Krishnamurthy
Princeton Physics

Declan Campbell
Princeton Neuroscience

Yuang Yang
PrincetonElectrical Engineering

Awni Altabaa
Yale Computer Science

Funding

Office of Naval Research

Vannevar Bush Faculty Fellowship

Recent Works

Emergent Symbols Through Binding Network (ESBN)
(Webb et al., ICML 2021)

The Relational Bottleneck: An Inductive Bias for Abstraction
(Webb et al., TICS 2023)

Abstractors: Transformers with Relational Cross Attention
(Altabaa et al., ICLR, 2024)

Slot Abstractors: Scalable Abstract Visual Reasoning
(Mondal et al., NeurIPS 2023)

Symbolic Abstraction Heads in LLMs
(Yang et al., arXiv:2502.20332 2025)

Limits to Vision Language Models and Compositional Coding
(Campbell et al., NeurIPS, 2024)

Episodic Generalization and Optimization
(Giallanza et al., OpenMind, 2024)