S&DS 365 / 665

Intermediate Machine Learning

Transformers, Abstraction, Reasoning, and AI Safety

December 4

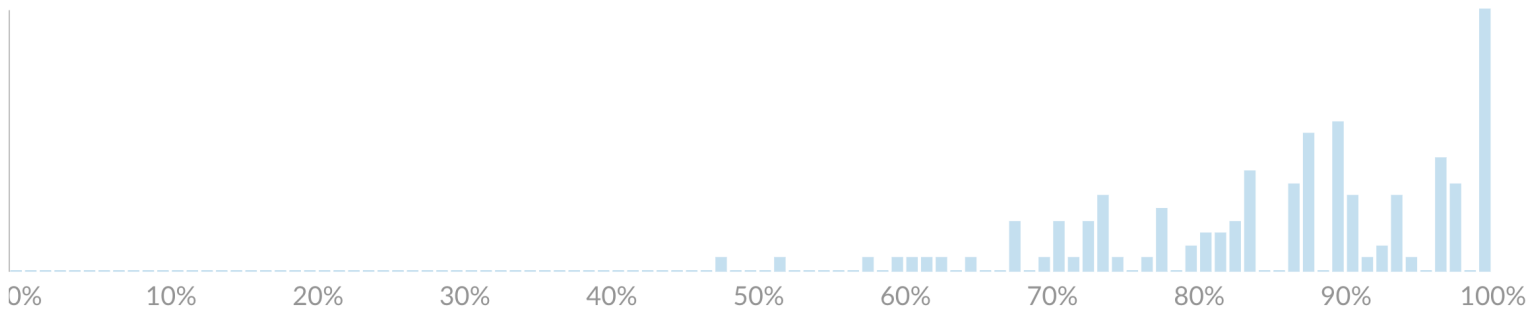
Yale

Reminders

- Quiz 5 last week
- Assn 5 out; due this Wednesday
- Final exam: Wednesday Dec 20, 9am in HQ L02
- Practice exams are posted

Quiz 5

Average Score	High Score	Low Score	Standard Deviation	Average Time
86%	100%	48%	1.15	17:18



For Today

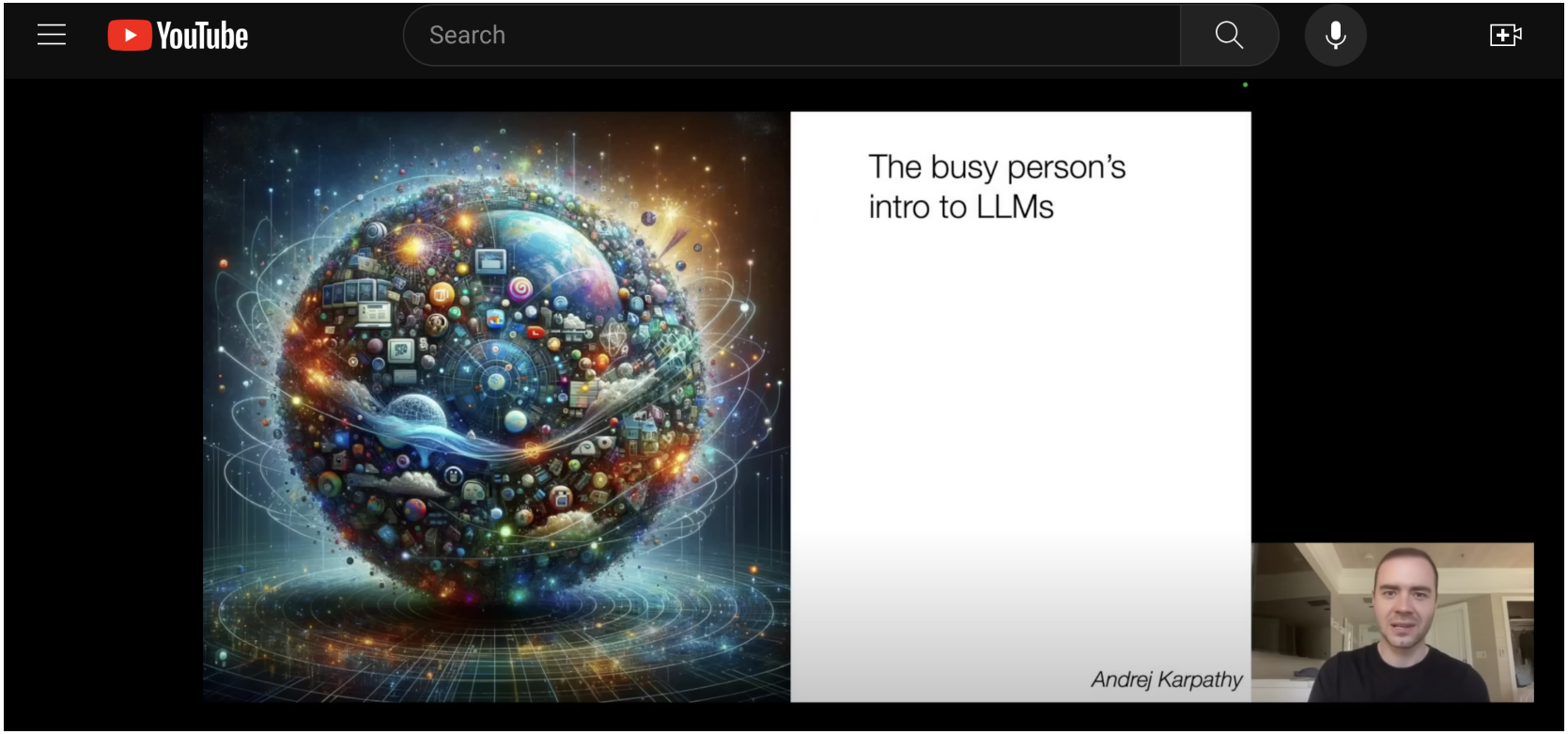
- An insider's view on large language models
  - LLMs as compressors
  - Scaling laws
  - Finetuning
  - Thinking slow
  - AI safety: attacks on LLMs

Karpathy's video is excellent (IMHO). Please watch at least the segments linked to; I hope you might enjoy the entire video.

For Today

- An insider's view on large language models
- Abstraction and reasoning: The next frontier?
  - Some recent research
  - Not required

LLMs: A recent, expert overview



https://www.youtube.com/watch?v=zjkBMFhNj_g

Posted November 22, 2023

I'm indebted to Andrej Karpathy—a master explainer/teacher in a corporate world.

LLMs are Internet Compressors

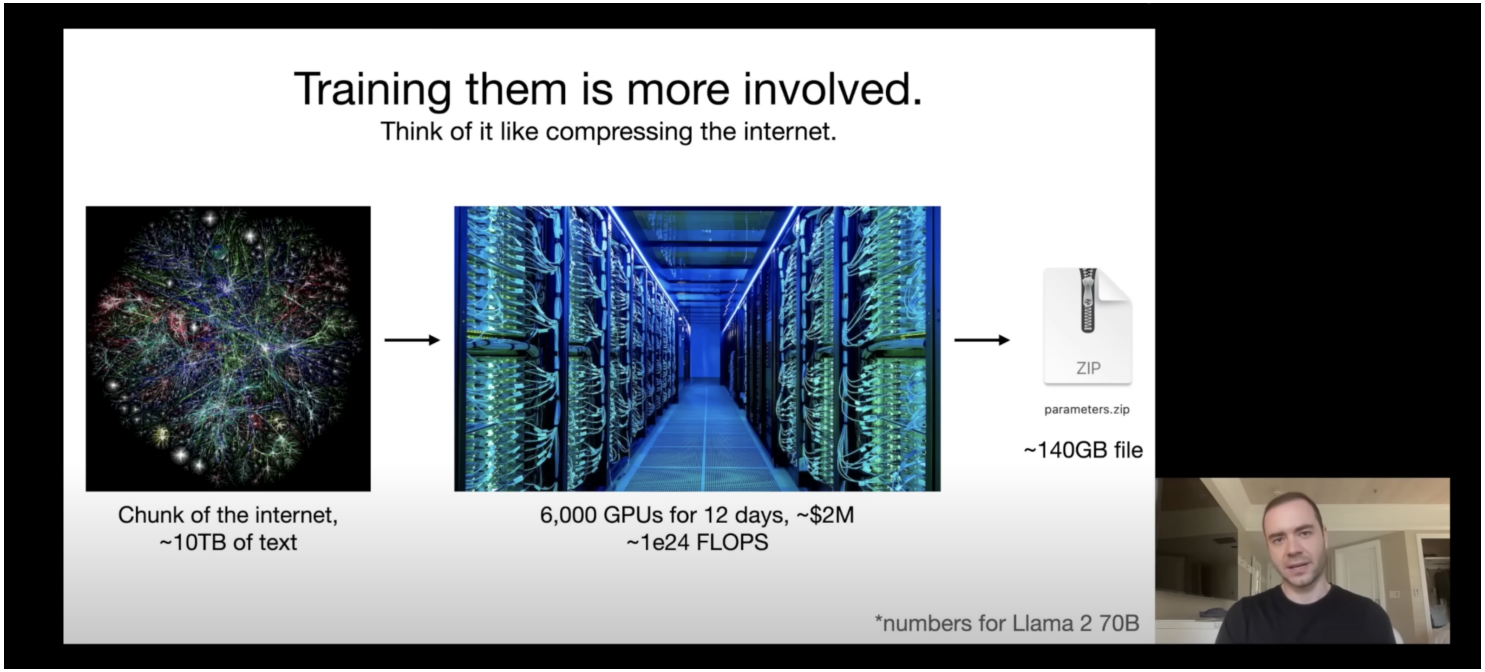
LLMs are Internet Compressors

- If you can predict the next symbol well, you can compress well
- The technical basis is *arithmetic coding* in information theory
- *The parameters in a LLM transformer are a compressed version of the internet*

---

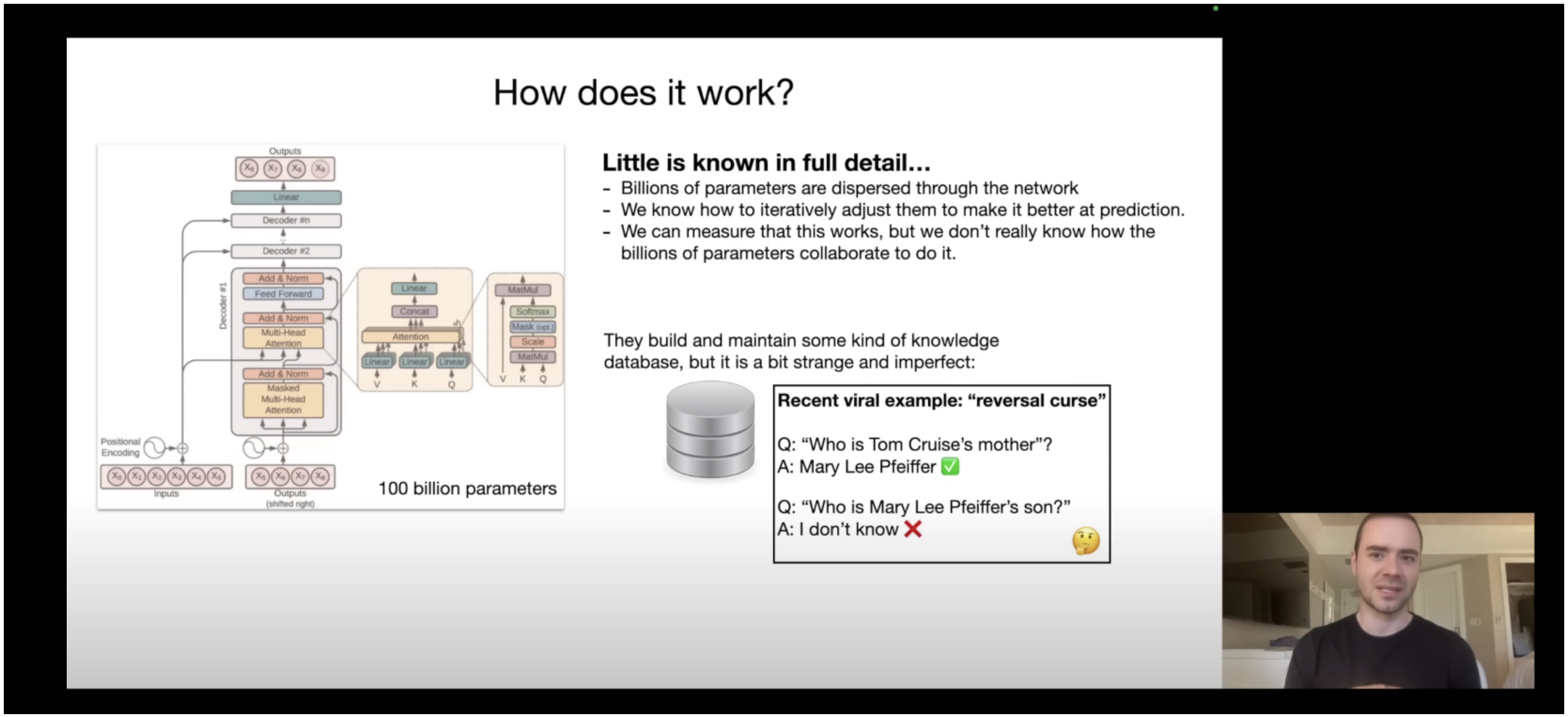
https://en.wikipedia.org/wiki/Arithmetic_coding

LLMs are Internet Compressors



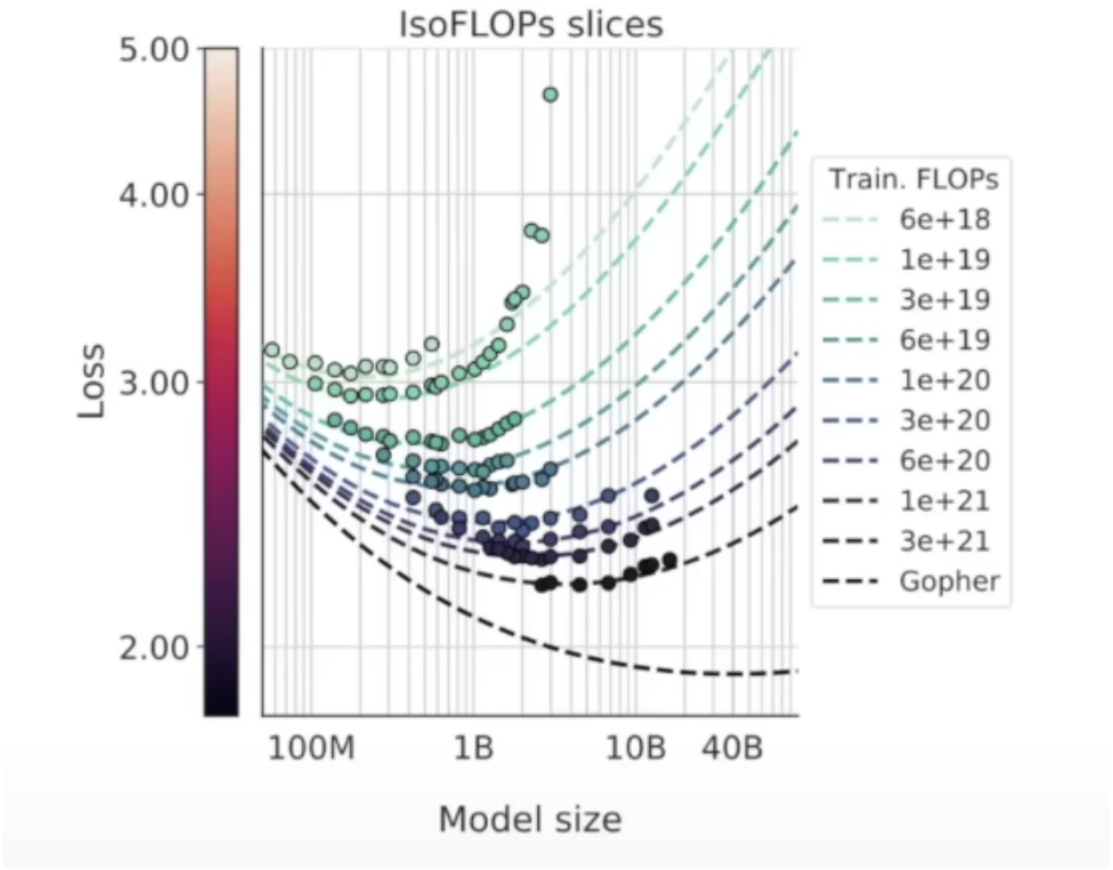
https://www.youtube.com/watch?v=zjkBMFhNj_g&t=4m10s

How does it work? We don't know



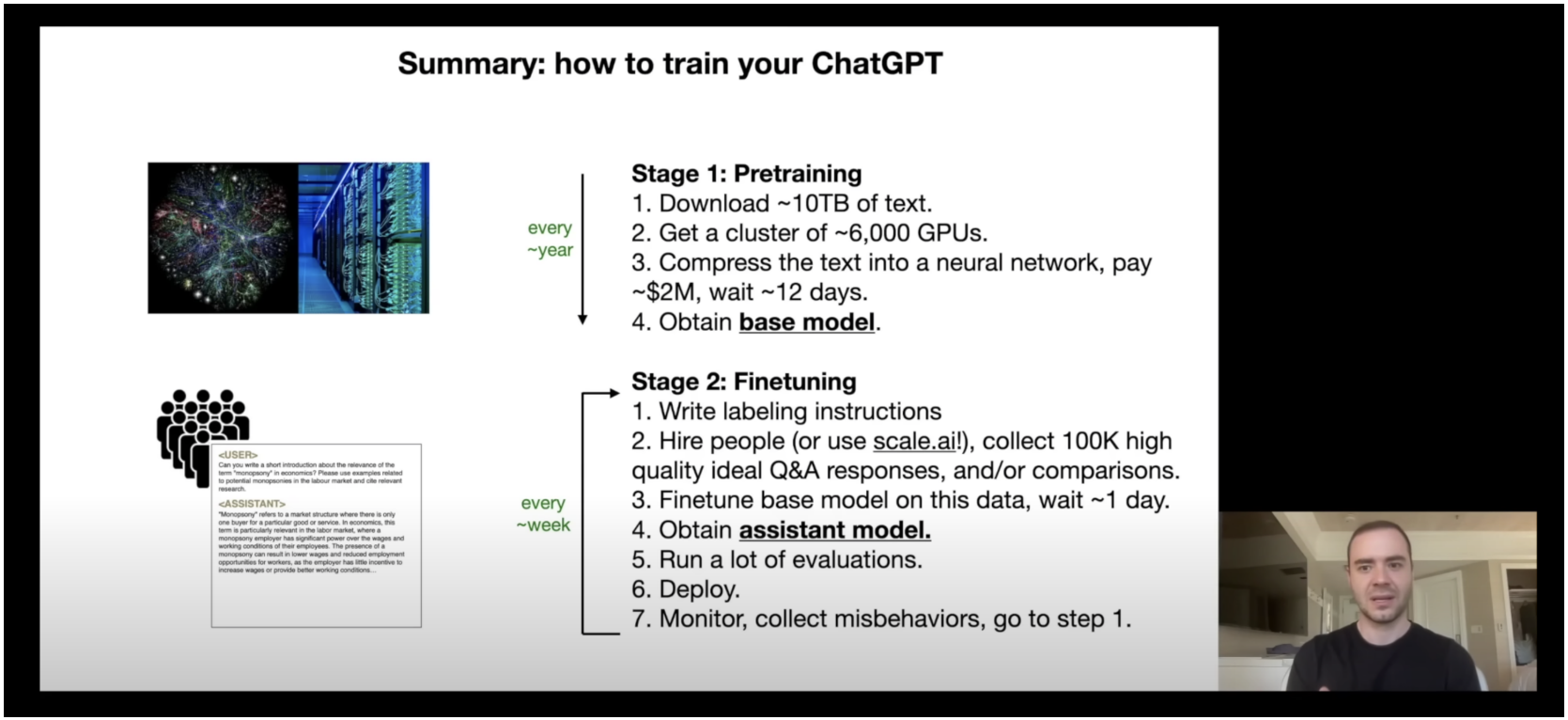
https://www.youtube.com/watch?v=zjkBMFhNj_g&t=11m25s

LLM scaling laws: Bigger is better



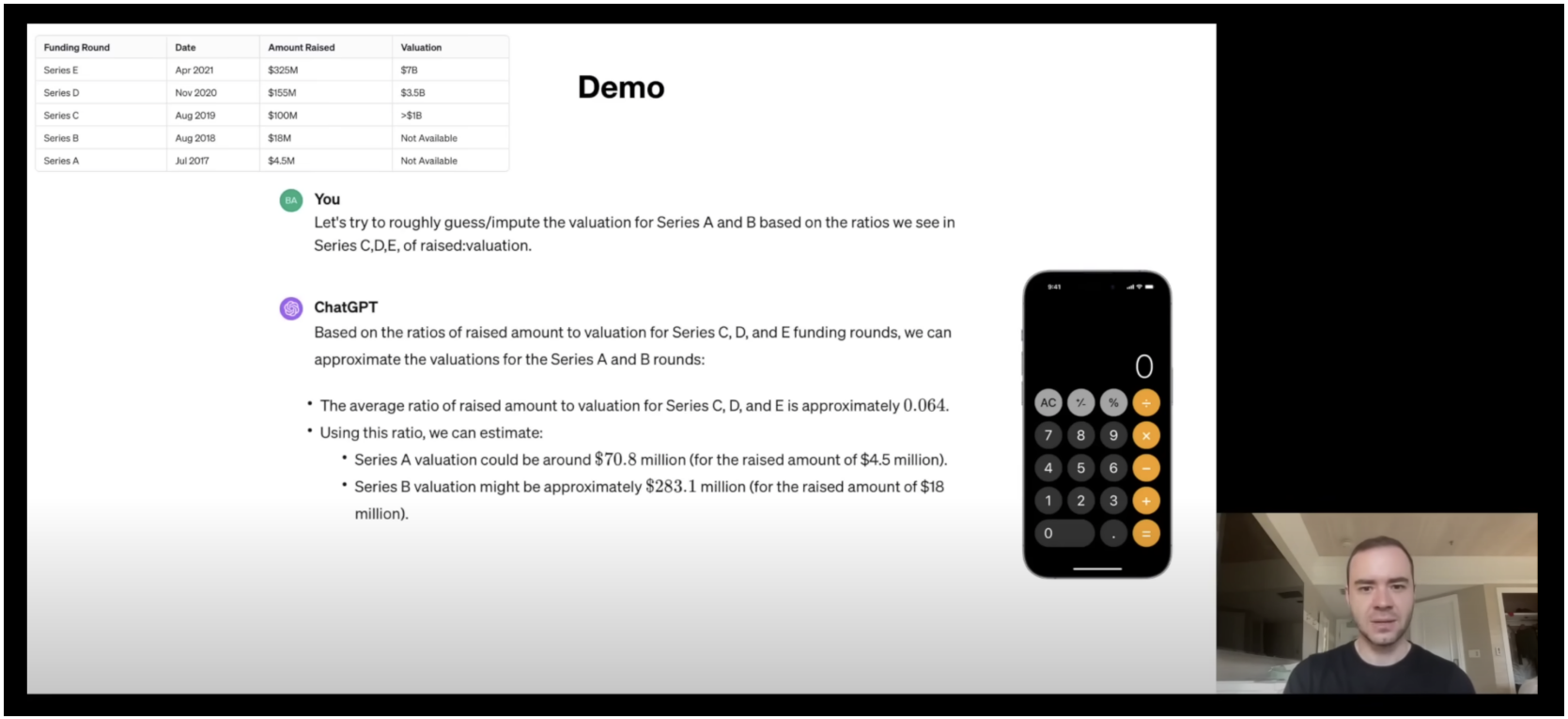
https://www.youtube.com/watch?v=zjkBMFhNj_g&t=25m40s

Finetuning



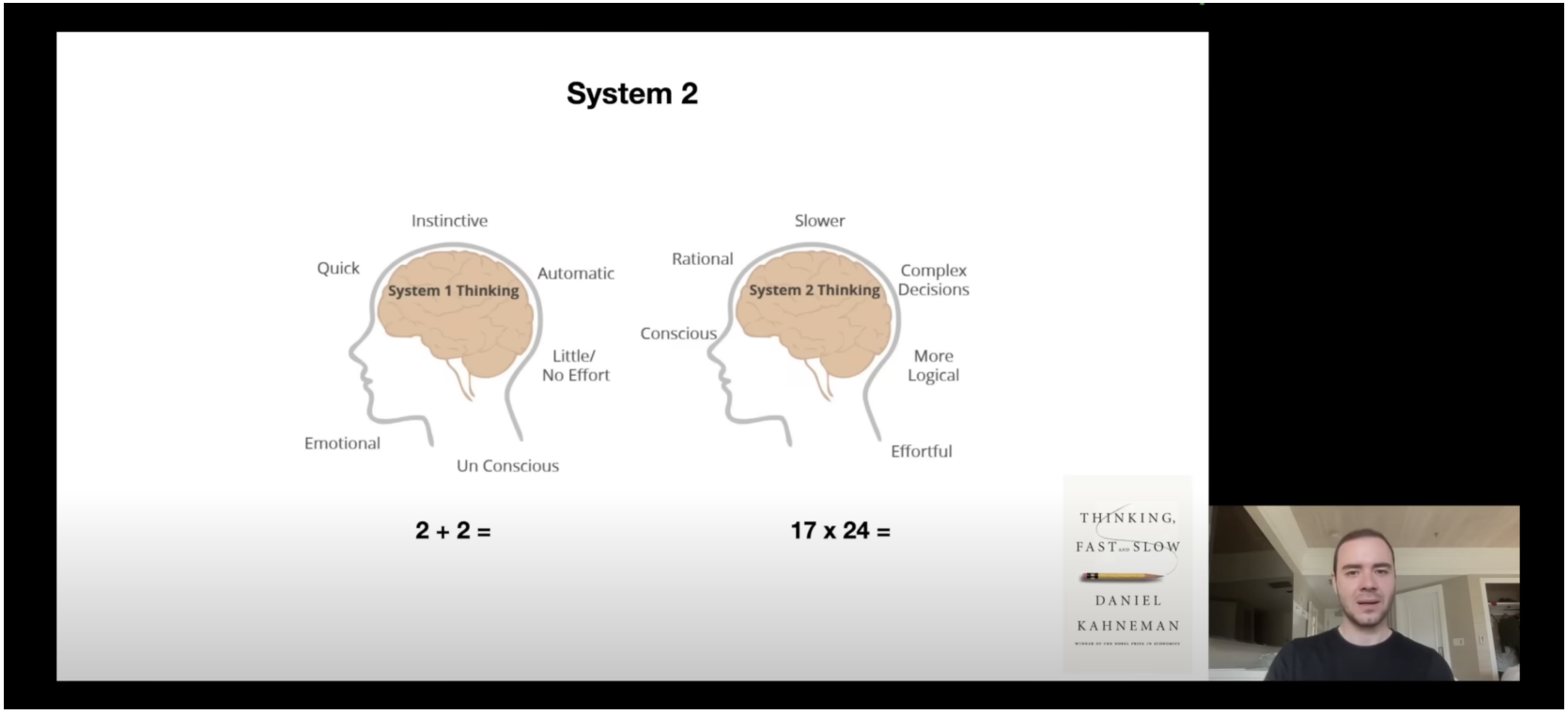
https://www.youtube.com/watch?v=zjkBMFhNj_g&t=14m19s

Advanced capabilities: Using tools



https://www.youtube.com/watch?v=zjkBMFhNj_g&t=27m44s

Thinking slower



https://www.youtube.com/watch?v=zjkBMFhNj_g&t=35m00s

Security

https://www.youtube.com/watch?v=zjkBMFhNj_g&t=45m43s

Jailbreaks: https://www.youtube.com/watch?v=zjkBMFhNj_g&t=46m15s

Prompt injection: https://www.youtube.com/watch?v=zjkBMFhNj_g&t=51m30s

Data poisoning: https://www.youtube.com/watch?v=zjkBMFhNj_g&t=56m23s

“We have Paleolithic emotions, medieval institutions and godlike technology. And it is terrifically dangerous, and it is now approaching a point of crisis overall.”

Edward O. Wilson (2009)

Summary: LLMs for busy people

- The billions of parameters can be thought of as a compressed version of the internet
- As LLMs are scaled up, we can expect further gains; they are not close to the prediction limit
- The models are well understood technically; why they actually work is not
- Finetuning is needed to make next word predictors useful
- The examples used for finetuning can employ tools like web browsers and calculators
- Reasoning (thinking “slower”) is a next frontier
- Fixing attacks on LLMs is a “cat and mouse game” that is a new type of computer security

Abstraction and reasoning with transformers

The next slides introduce some recent work in my own group, which is in the direction of Karpathy's discussion of "thinking slow"

This material is not required (that is, you will not be tested on it), but I hope you find it interesting and that it might solidify your understanding of this type of machine learning.

Reasoning in terms of relations, analogies, and abstraction is a hallmark of human intelligence and creativity

It is largely separate from function approximation for sensory tasks such as image and audio processing

How can abstract symbols emerge from distributed, neural representations?

Collaborators

Awni Altabaa
Yale

Jonathan Cohen
Princeton

Taylor Webb
UCLA

arXiv:2304.00195

Two types of intelligence

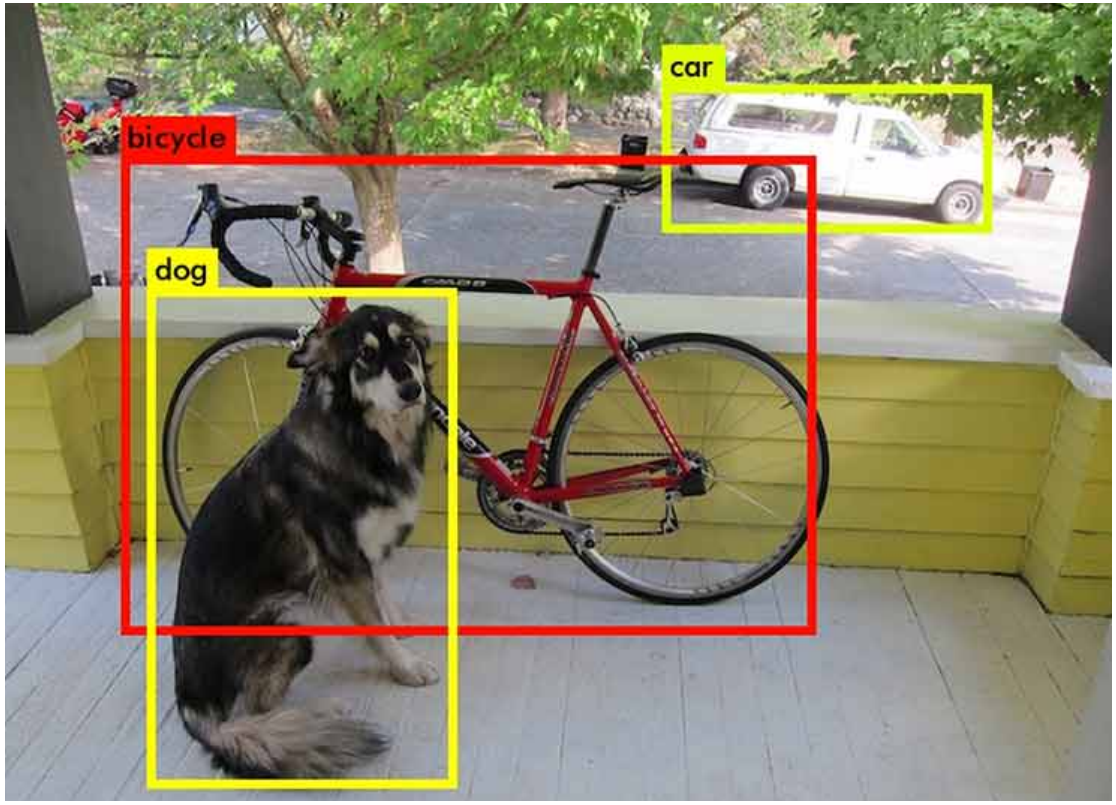
1. "Sensory/motor"— acquire semantic and procedural knowledge
   - Requires extensive data and training
   - Slow to learn, fast to apply
   - Well captured by modern deep learning

Two types of intelligence

1. “Sensory/motor”— acquire semantic and procedural knowledge

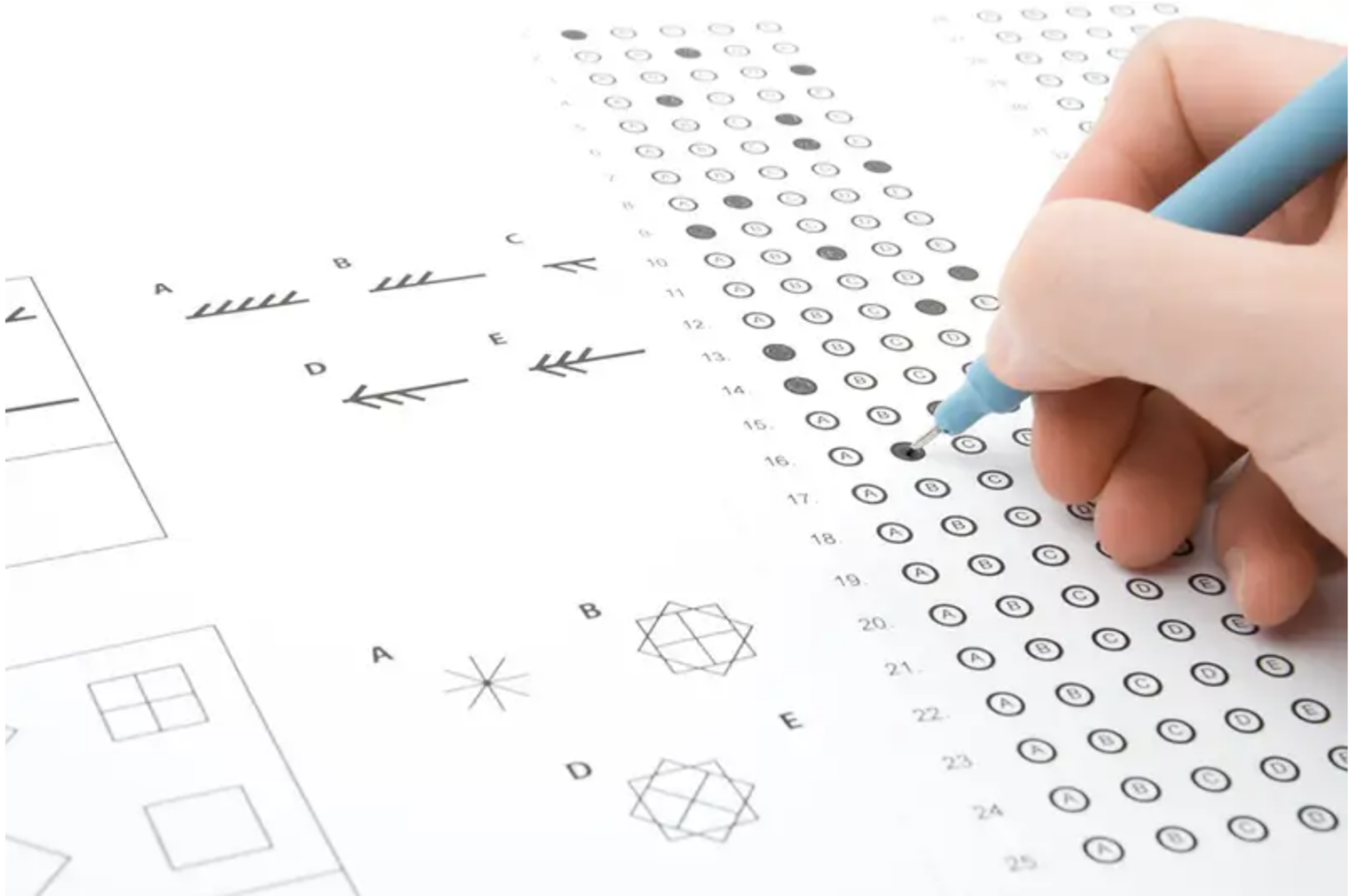
Two types of intelligence

1. “Sensory/motor”— acquire semantic and procedural knowledge

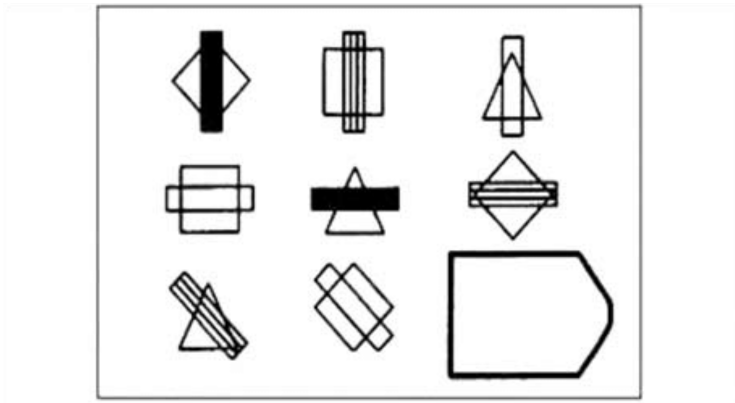


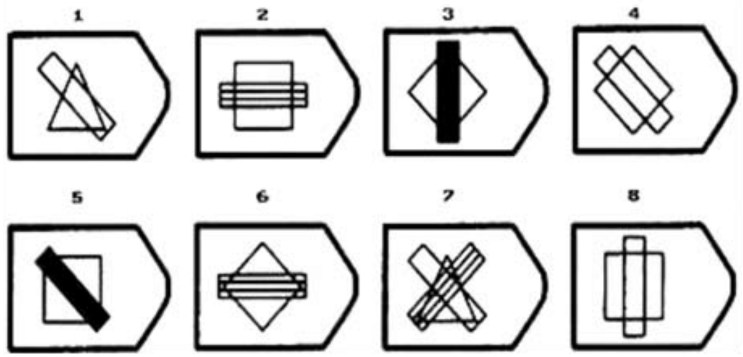
Two types of intelligence

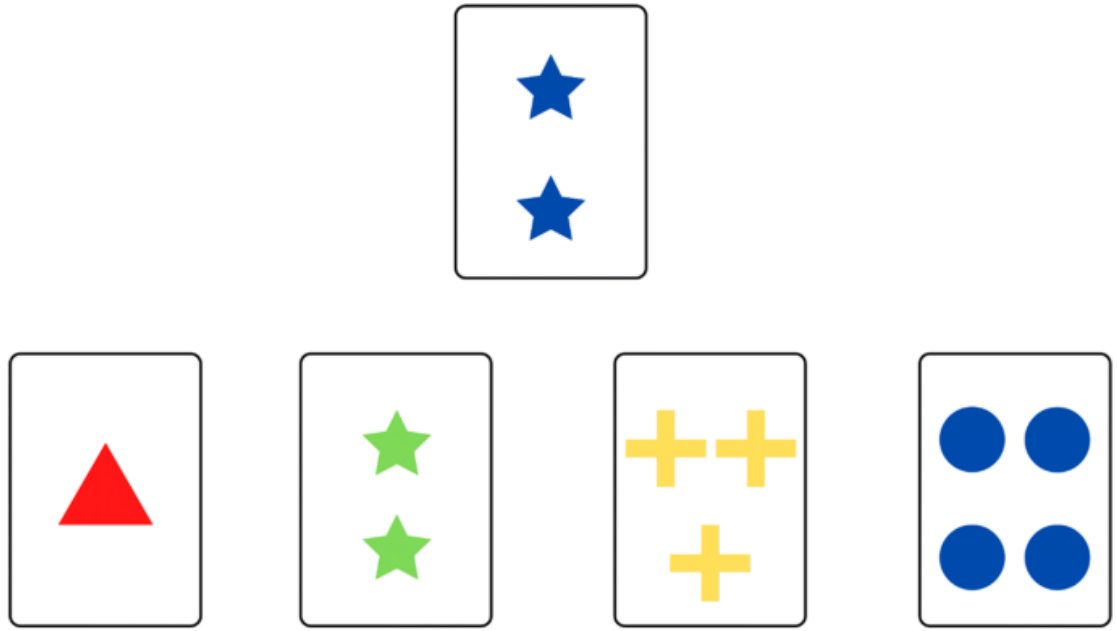
2. “Prefrontal”— identify novel associations and relations
   - Fast to learn, slow to apply
   - Symbolic processing and abstraction
   - Little explicit training data

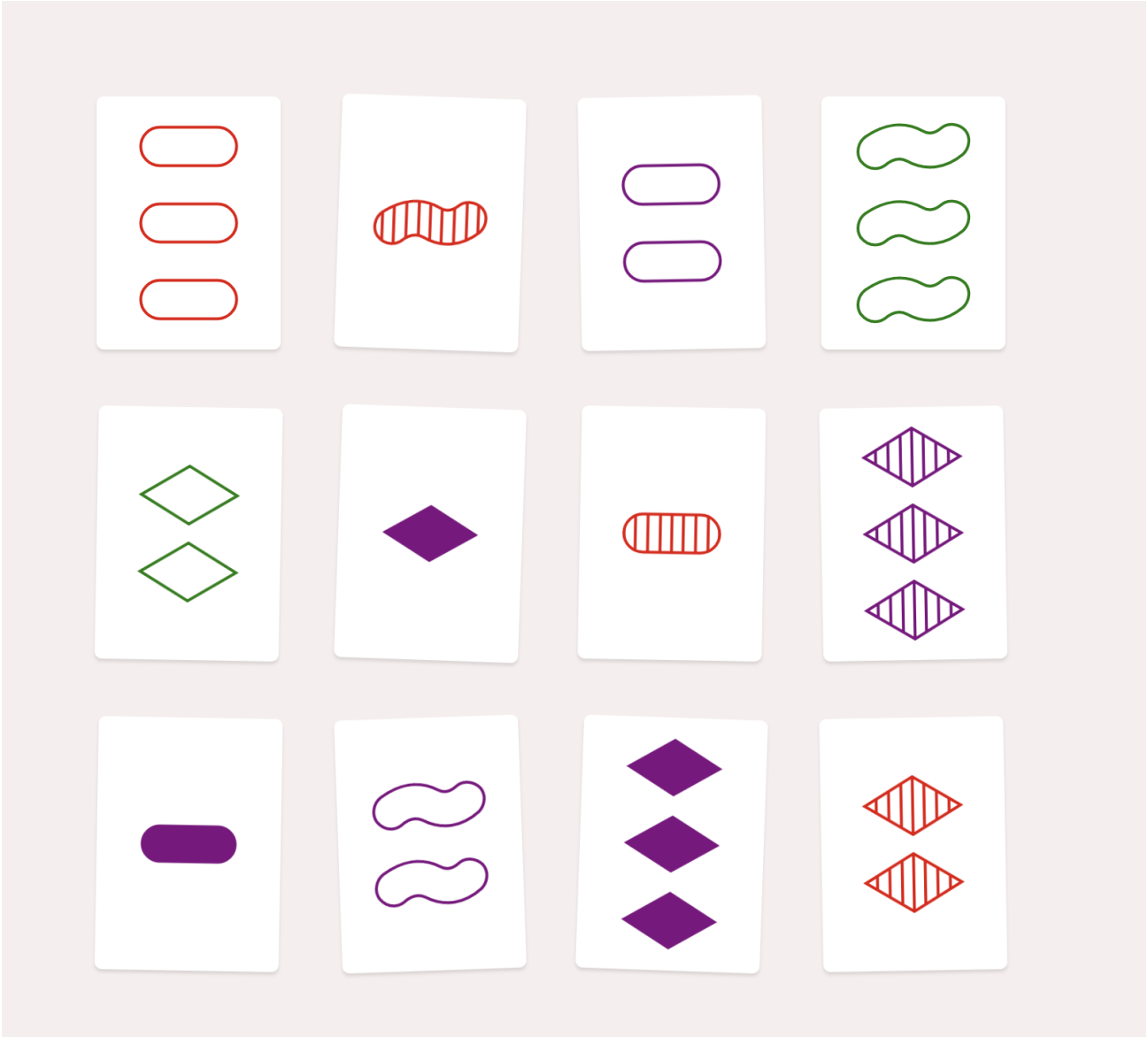


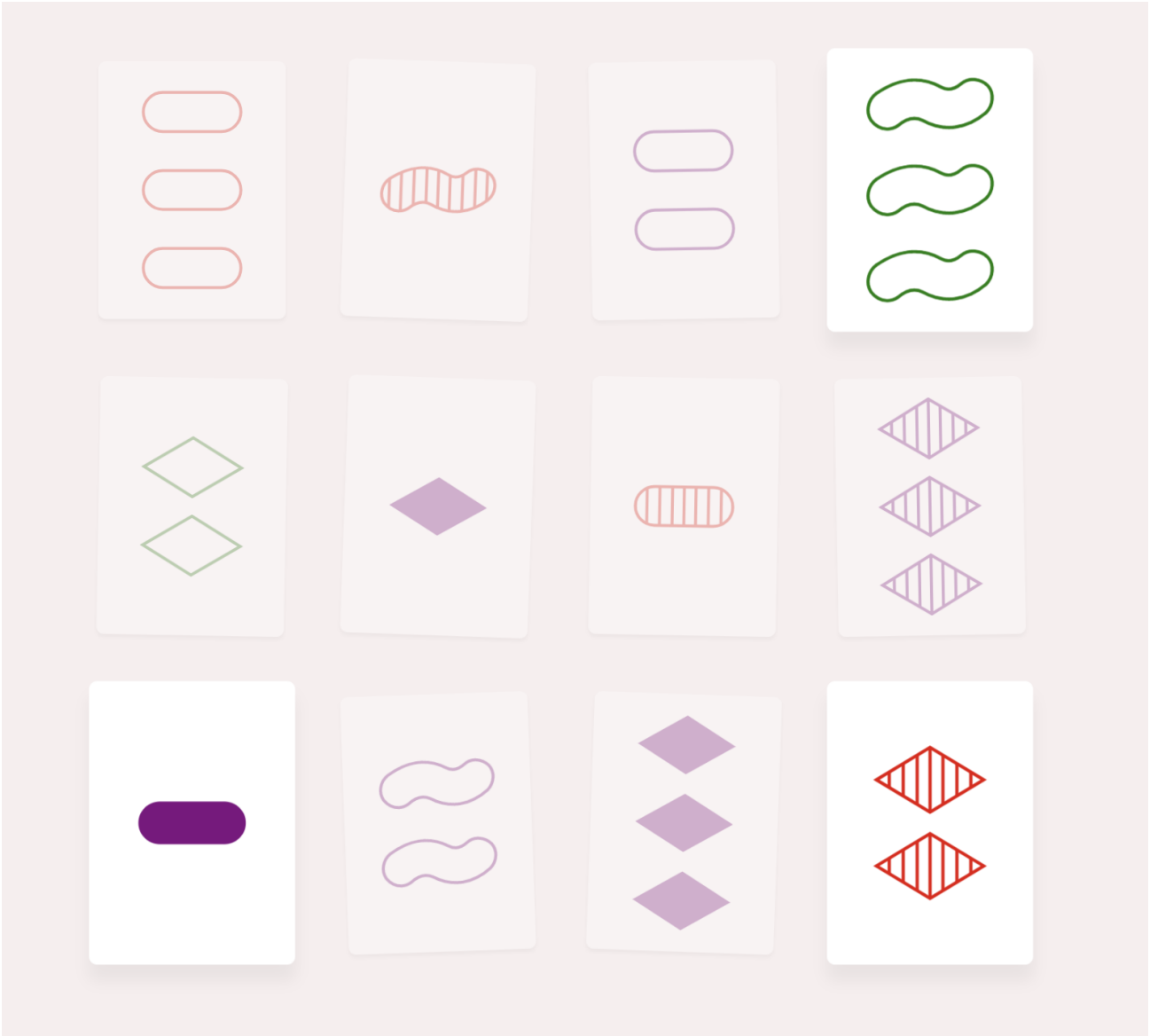
AI struggles with IQ tests
Panther Media GmbH / Alamy Stock Photo




1
2
3
4
5
6
7
8







Two types of intelligence

1. “Sensory/motor”— efficient function approximation
2. “Prefrontal”— rapid abstraction, association and generalization

Can both types be supported in a single architecture?

Relational Bottleneck

Key: An inductive bias we call the *relational bottleneck*:

Information flows from sensory and motor subsystems to reasoning and decision making subsystems through relations, as computed through inner products between distributed representations

---

Complementary Learning Systems theory of McClelland et al., 1995; suggestive of Crick's "Central Dogma."

Our work: Abstractors

- A framework that enforces the relational bottleneck
- Based on symbolic message passing and relation tensors
- Cast in terms of Transformer architectures

Previous work: “Neural Turing machines,” Graves et al., arXiv:1410.5401 ; “Emergent symbols through binding in external memory,” Webb et al., arXiv:2012.14601; “On Neural Architecture Inductive Biases for Relational Tasks,” Kerg et al., arXiv:2206.05056

Relations are essential to reasoning

“Jack and Jill are sitting side by side. The person next to Jack is angry. The person next to Jill is happy. Who is happy, Jack or Jill?”

Shlomi Sher, “On Artifice and Intelligence: How to spot counterfeit cognition,” medium.com/@shlomi.sher, April 21, 2023

Relations are essential to reasoning

ChatGPT response (yesterday):

Based on the information provided, Jill is happy. The statement specifies that the person next to Jill is happy, indicating that Jill is the one who is happy. Therefore, Jill is happy while Jack is angry.

Relations are essential to reasoning

ChatGPT response (yesterday):

Based on the information provided, Jill is happy. The statement specifies that the person next to Jill is happy, indicating that Jill is the one who is happy. Therefore, Jill is happy while Jack is angry.

Relations are essential to reasoning

ChatGPT response (yesterday):

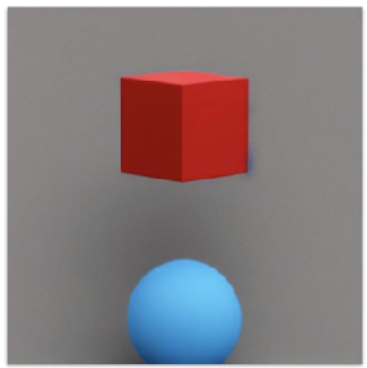
> Based on the information provided, Jill is happy. The statement specifies that the person next to Jill is happy, indicating that Jill is the one who is happy. Therefore, Jill is happy while Jack is angry.

Relations are essential to reasoning

“Two shapes are in a room. The shape next to the cube is blue. The shape next to the sphere is red. Which is red, the cube or the sphere?”

Relations are essential to reasoning

"Two shapes are in a room. The shape next to the cube is blue. The shape next to the sphere is red. Which is red, the cube or the sphere?"



Relations are essential to reasoning

Based on the given information, it can be inferred that the sphere is red. The statement specifies that the shape next to the sphere is red, implying that the sphere itself is red. The color of the cube is not mentioned, so we cannot determine its color from the given information.

Relations are essential to reasoning

Based on the given information, it can be inferred that the sphere is red. The statement specifies that the shape next to the sphere is red, implying that the sphere itself is red. The color of the cube is not mentioned, so we cannot determine its color from the given information.

Relations are essential to reasoning

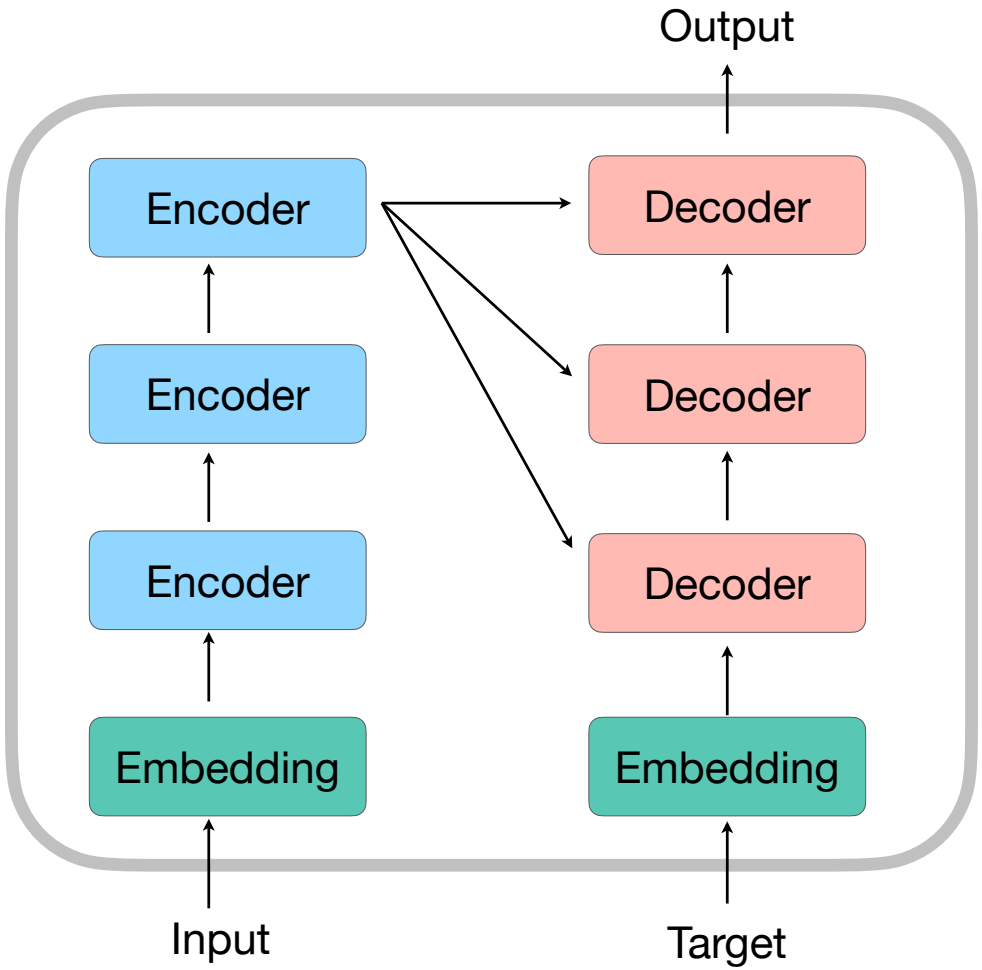
Based on the given information, it can be inferred that the sphere is red. The statement specifies that the shape next to the sphere is red, implying that the sphere itself is red. The color of the cube is not mentioned, so we cannot determine its color from the given information.

Transformer architecture

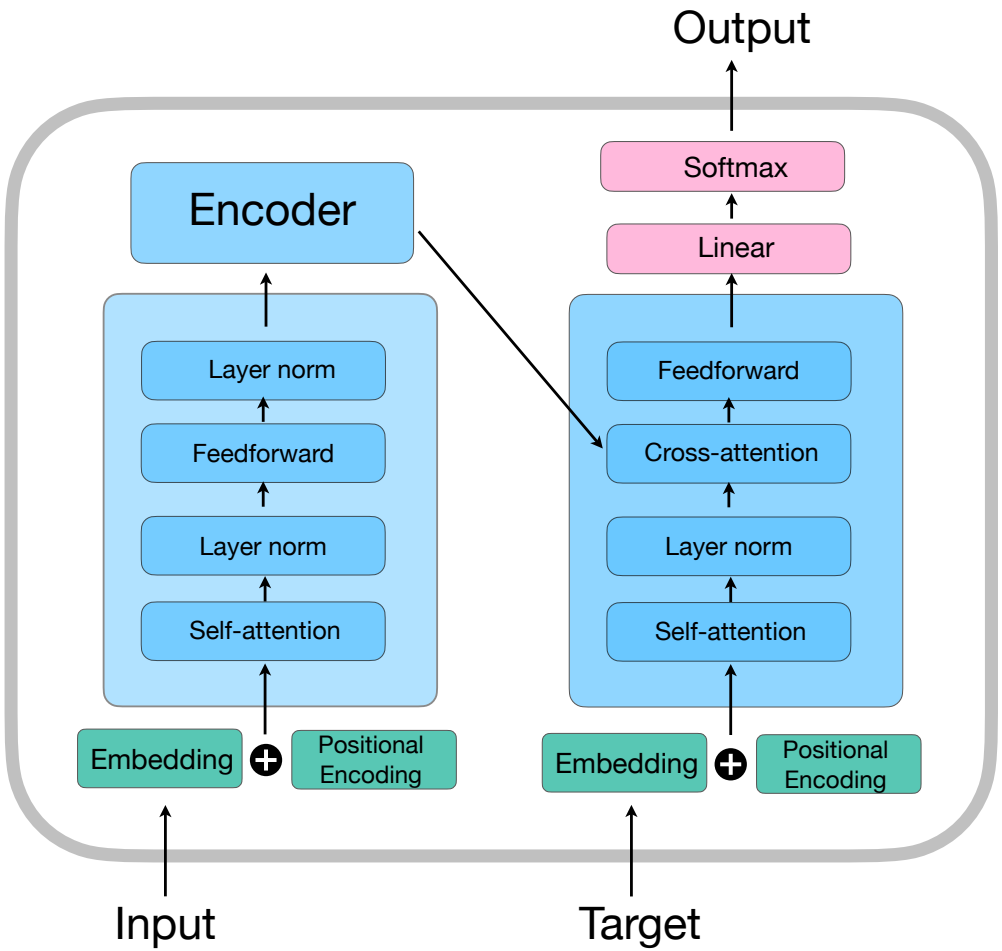
A Transformer is a seq2seq model based on encoder and decoder modules.

Transformers are powerful alternatives to RNNs that transform the encoder/decoder states using (multi-head) attention mechanisms.

Transformer architecture



Transformer architecture



Two encoder layers and one decoder layer

The two elephants played with an orange ball

The two elephants **played** with an orange **ball**

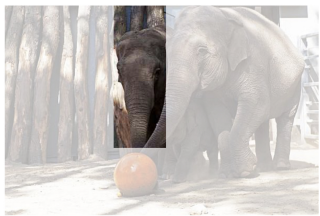
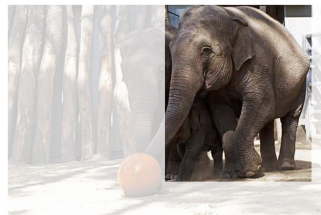
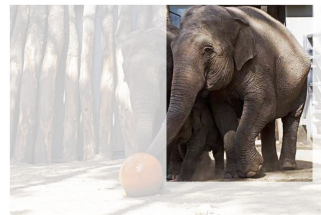
Self-attention: played → ball

Cross-attention:  → ball

Attention mechanisms

Self-attention:

$$E \longleftarrow \text{Attention}\,(Q \leftarrow E, K \leftarrow E, V \leftarrow E)$$

$Q \leftarrow$  $K \leftarrow$  $V \leftarrow$ 

Attention mechanisms

Self-attention:

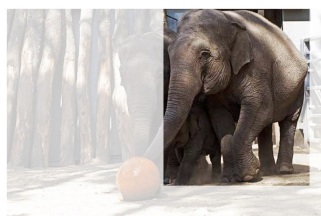
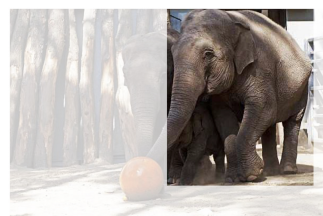
$$D \longleftarrow \text{Attention}\,(Q \leftarrow D, K \leftarrow D, V \leftarrow D)$$

$$Q \leftarrow \texttt{played} \quad K \leftarrow \texttt{ball} \quad V \leftarrow \texttt{ball}$$

Attention mechanisms

Cross-attention:

$$D \longleftarrow \text{Attention}\,(Q \leftarrow D, K \leftarrow E, V \leftarrow E)$$

$Q \leftarrow$ `elephants` $K \leftarrow$  $V \leftarrow$ 

Relational symbolic message passing

Symbolic message passing uses learnable, input-indepedent vectors $S_1, \ldots, S_m \in \mathbb{R}^d$ we call *symbols*.

Symbols are sent as messages on graphs with edge weights determined by relations between sensory inputs.

$$A_i \leftarrow \sum_j R(E_i, E_j) S_j$$

This is how abstraction is achieved.

Relational symbolic message passing

Symbolic message passing uses learnable, input-indepedent vectors $S_1, \ldots, S_m \in \mathbb{R}^d$ we call *symbols*.

Symbols are sent as messages on graphs with edge weights determined by relations between sensory inputs.

$$A_i \leftarrow \sum_j R(E_i, E_j) S_j$$

This is how abstraction is achieved.

Relational symbolic message passing

Symbolic message passing uses learnable, input-indepedent vectors $S_1, \ldots, S_m \in \mathbb{R}^d$ we call *symbols*.

Symbols are sent as messages on graphs with edge weights determined by relations between sensory inputs.

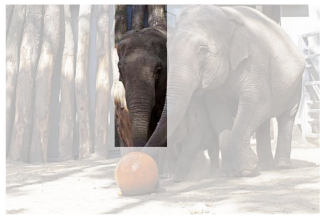
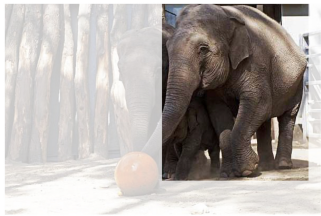
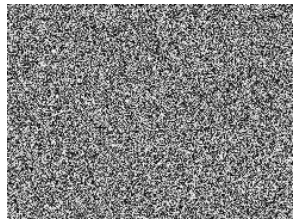
$$A_i \leftarrow \sum_j R(E_i, E_j) S_j$$

This is how abstraction is achieved.

Attention mechanisms

Relational Cross-attention:

$$A \longleftarrow \textbf{Attention}\,(Q \leftarrow E, K \leftarrow E, V \leftarrow A)$$

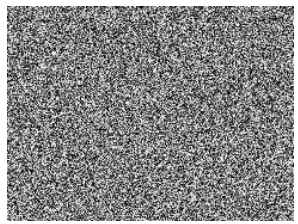
$Q \leftarrow$  $K \leftarrow$  $V \leftarrow$ 

Enforces relational bottlneck to allow abstraction and generalization

Attention mechanisms

Relational Cross-attention:

$$\textbf{Attention}\,(Q \leftarrow E, K \leftarrow E, V \leftarrow S)$$

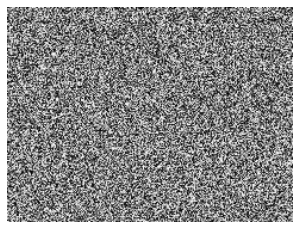
$Q \leftarrow$  $K \leftarrow$  $V \leftarrow$ 

Enforces relational bottlneck to allow abstraction and generalization

Attention mechanisms

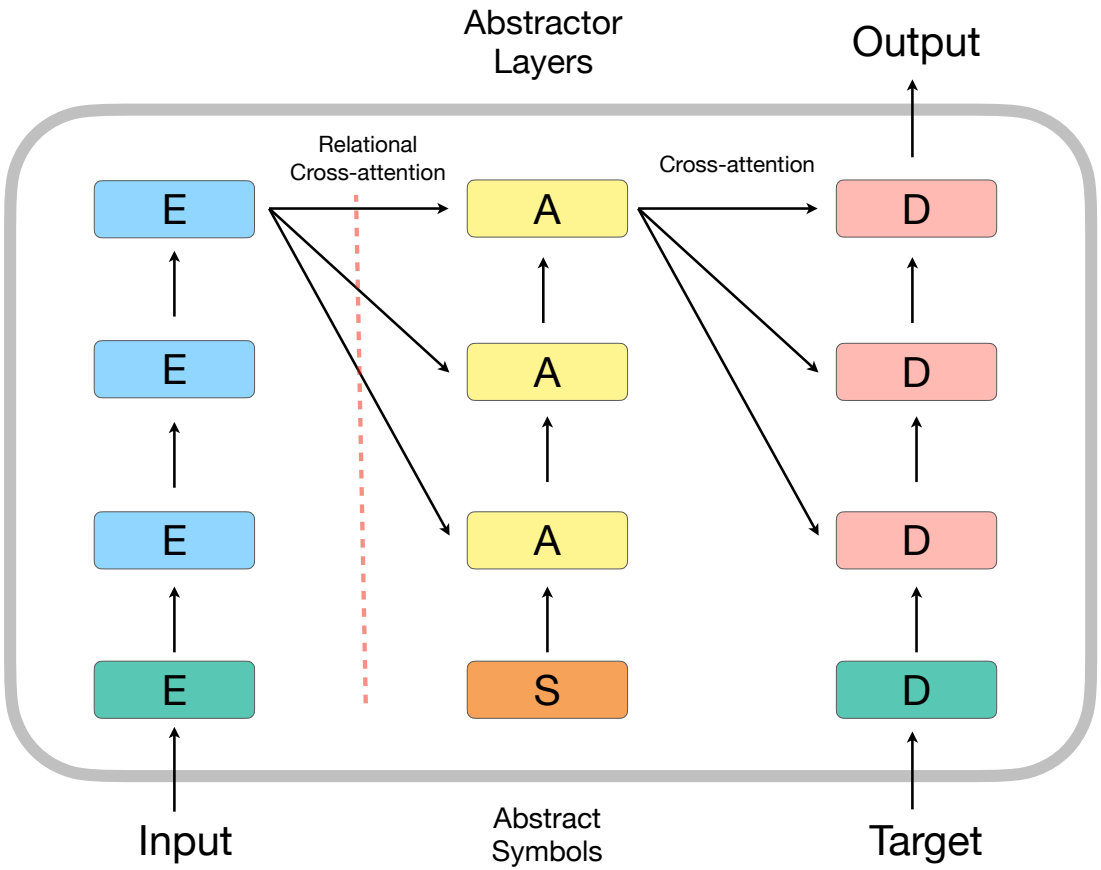
Relational Cross-attention:

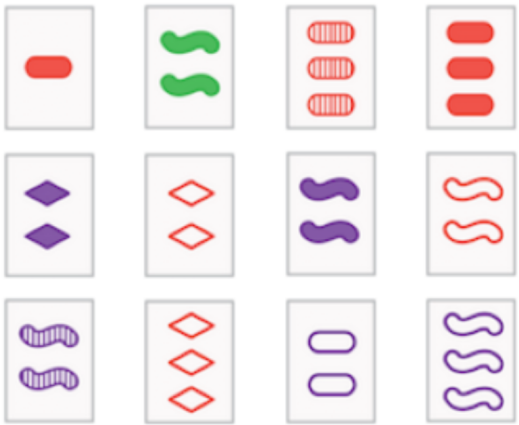
$$\textbf{Attention}\,(Q \leftarrow E, K \leftarrow E, V \leftarrow S)$$

$Q \leftarrow$  $K \leftarrow$  $V \leftarrow$ 

Enforces relational bottlneck to allow abstraction and generalization

Abstractor architecture



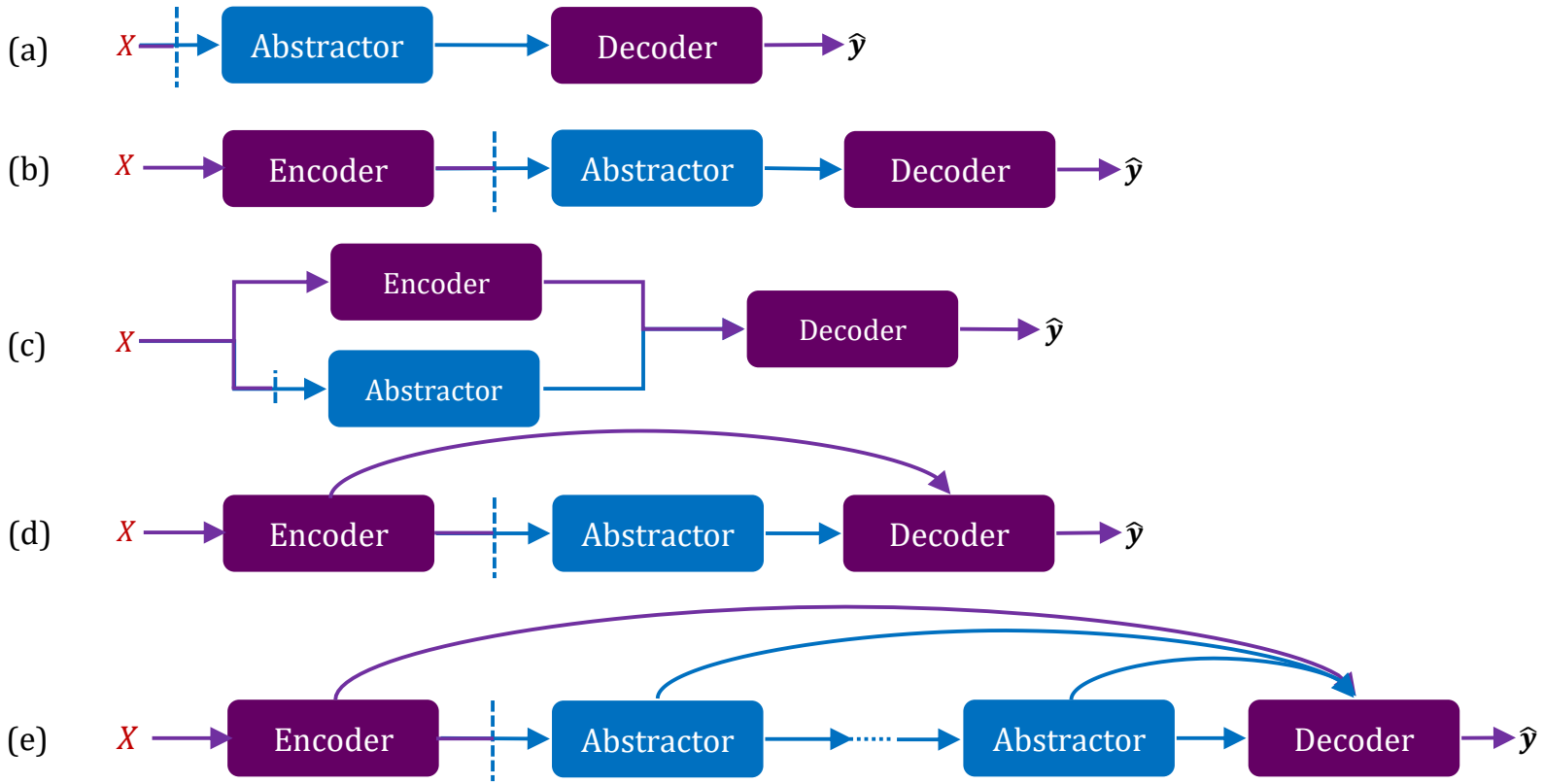


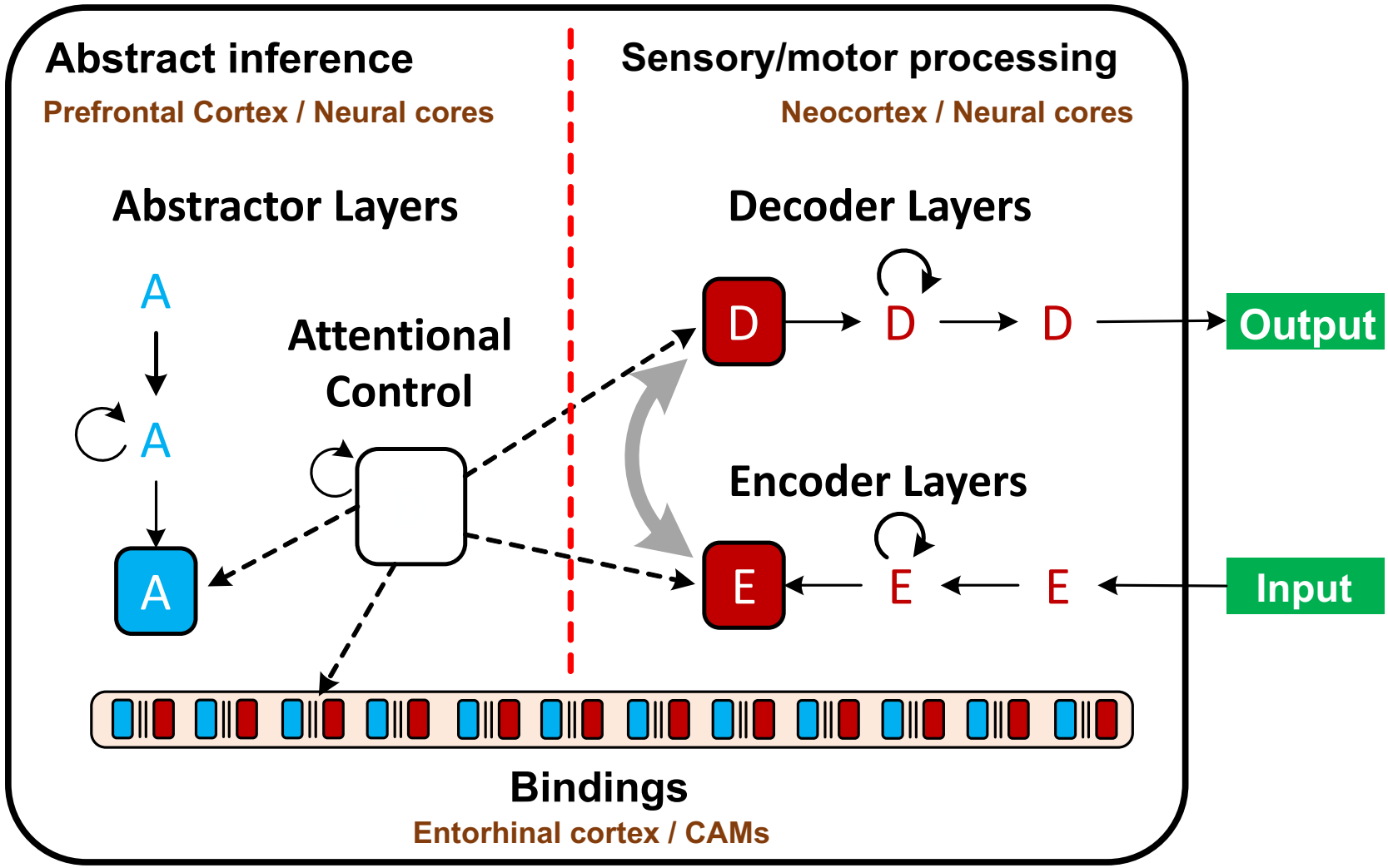
Attributes of encoder state E

R R R	R B G	R R B
$\frac{1}{3}(S_1 + S_2 + S_3)$	S_1	$\frac{1}{2}(S_1 + S_2)$
$\frac{1}{3}(S_1 + S_2 + S_3)$	S_2	$\frac{1}{2}(S_1 + S_2)$
$\frac{1}{3}(S_1 + S_2 + S_3)$	S_3	S_3

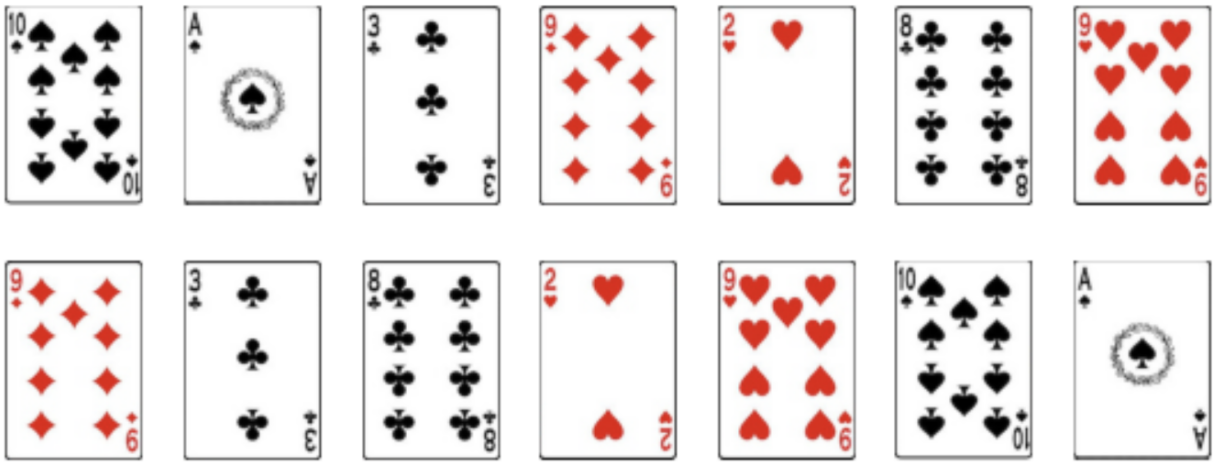
Transformed abstract symbol A

Abstractor architectures

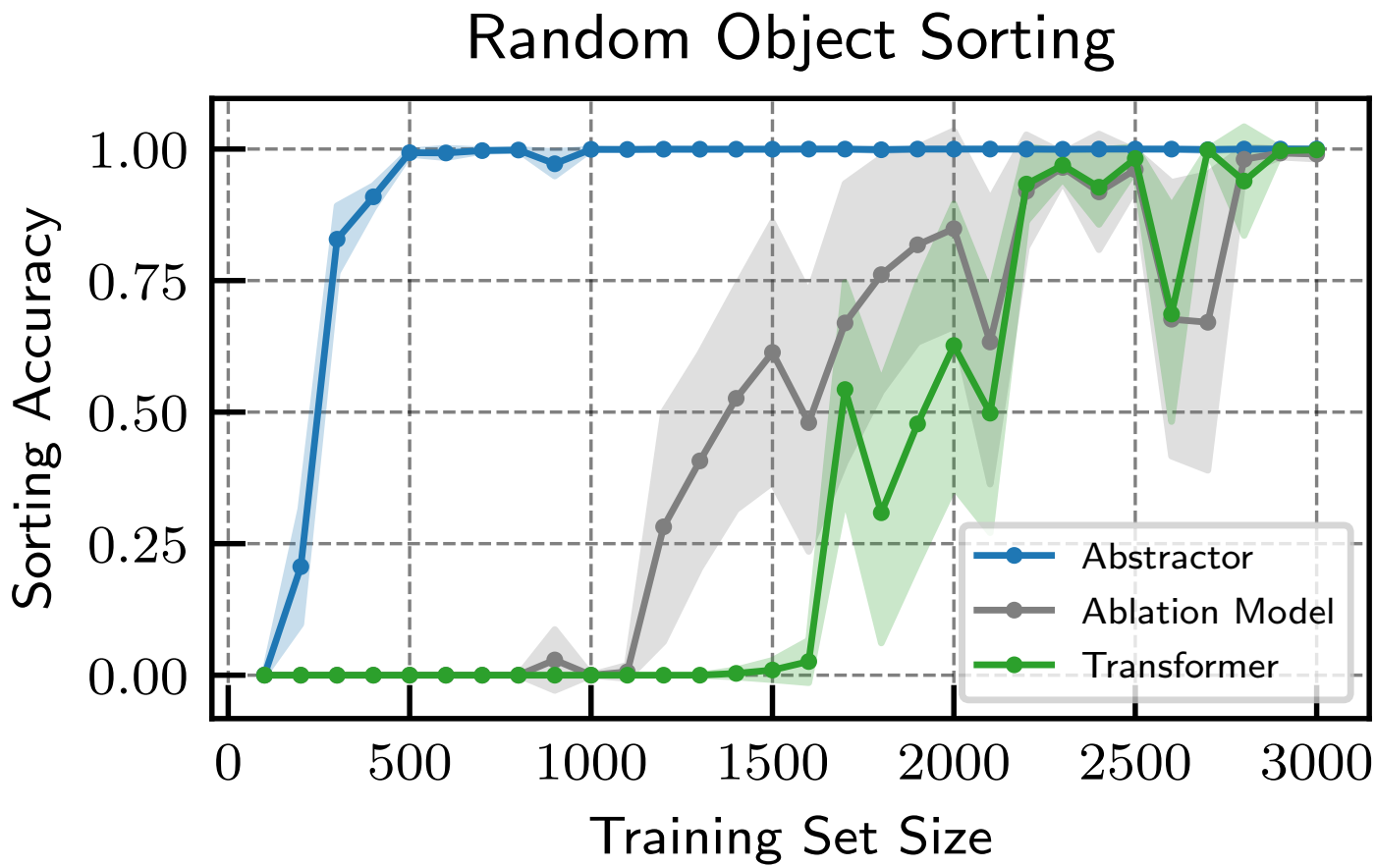



Abstract inference
Prefrontal Cortex / Neural cores
Sensory/motor processing
Neocortex / Neural cores
Abstractor Layers
Decoder Layers
A
A
A
Attentional
Control
D
D
D
Output
Encoder Layers
E
E
E
Input
Bindings
Entorhinal cortex / CAMs

Sorting

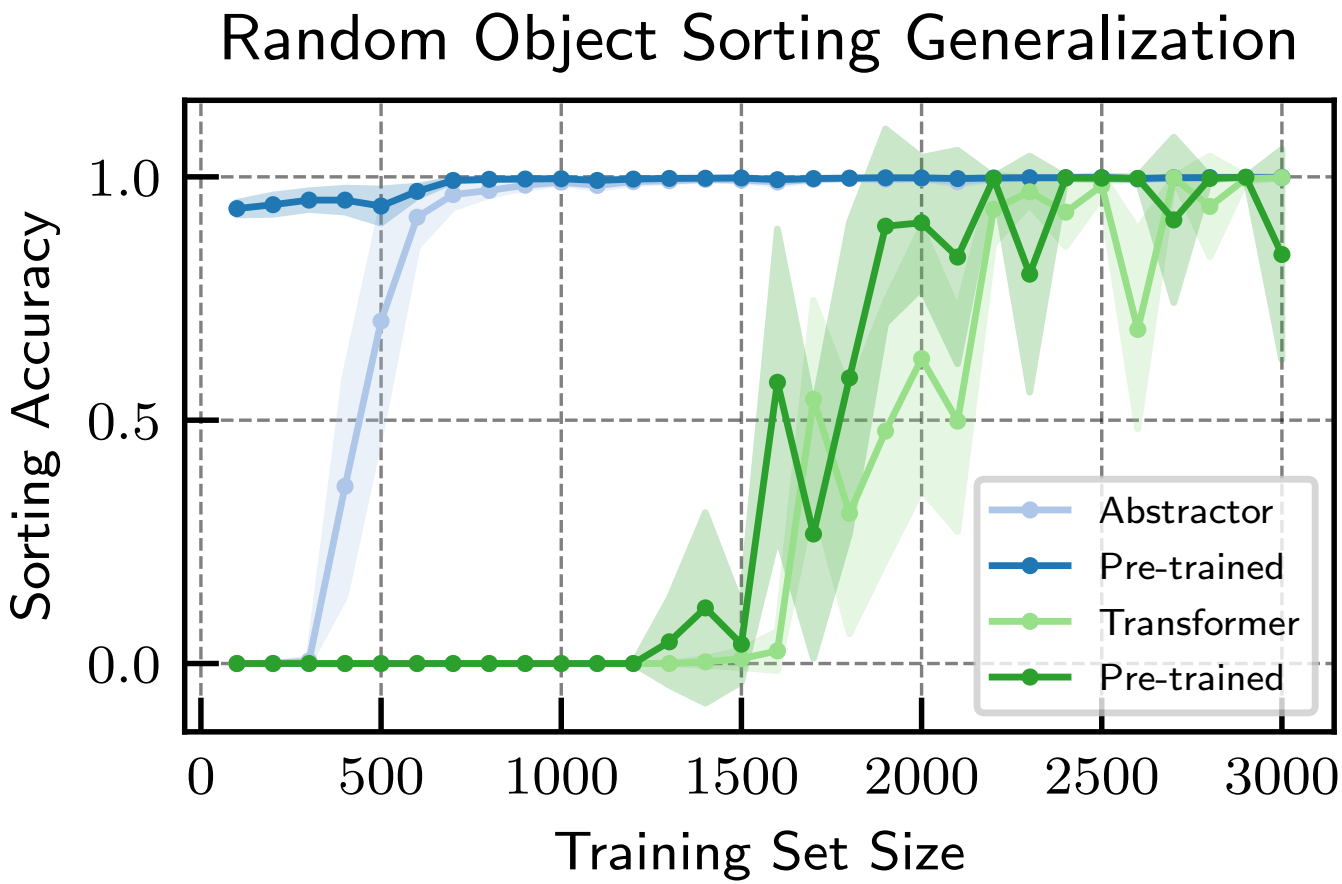


Sample efficiency and abstraction

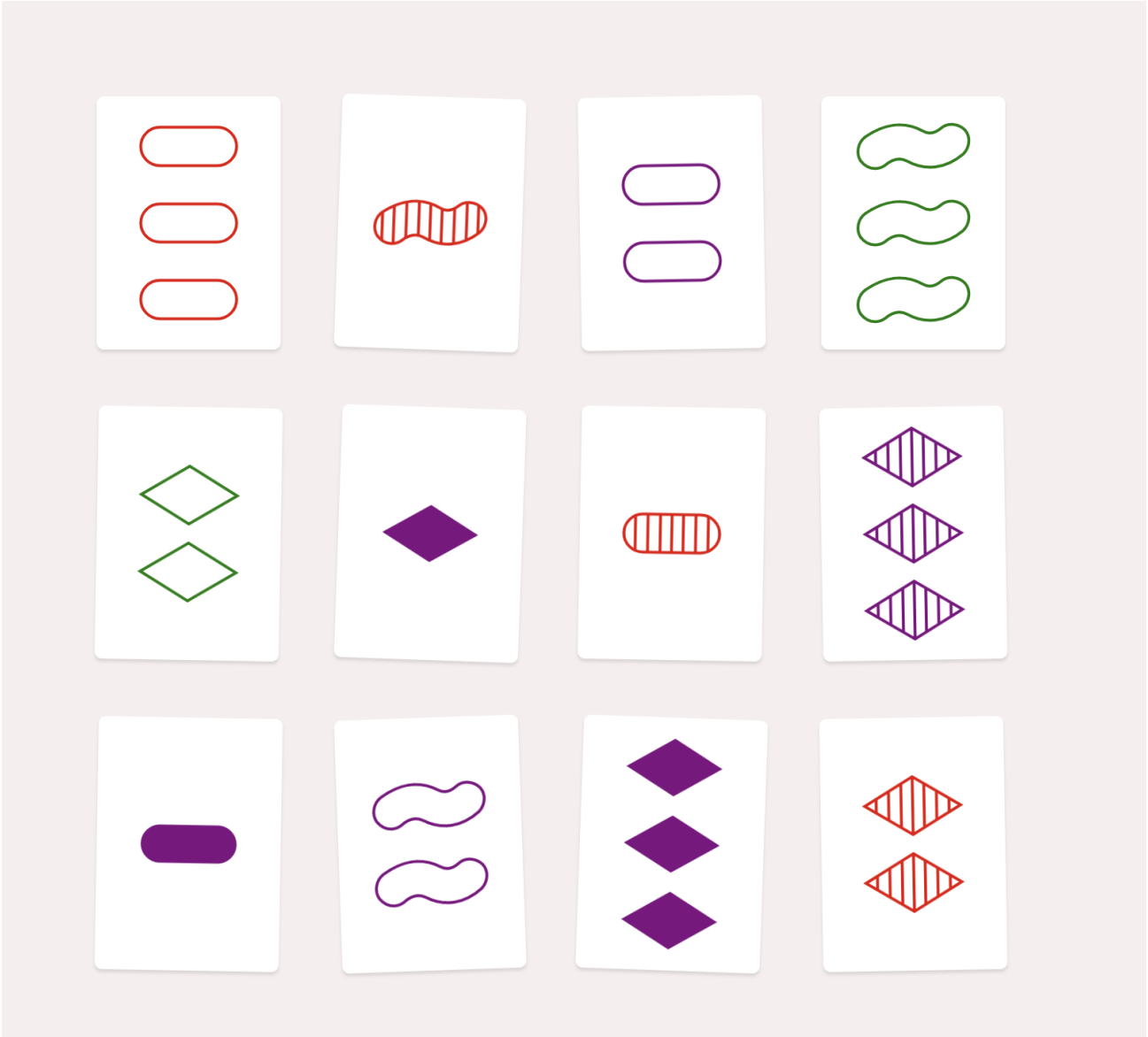


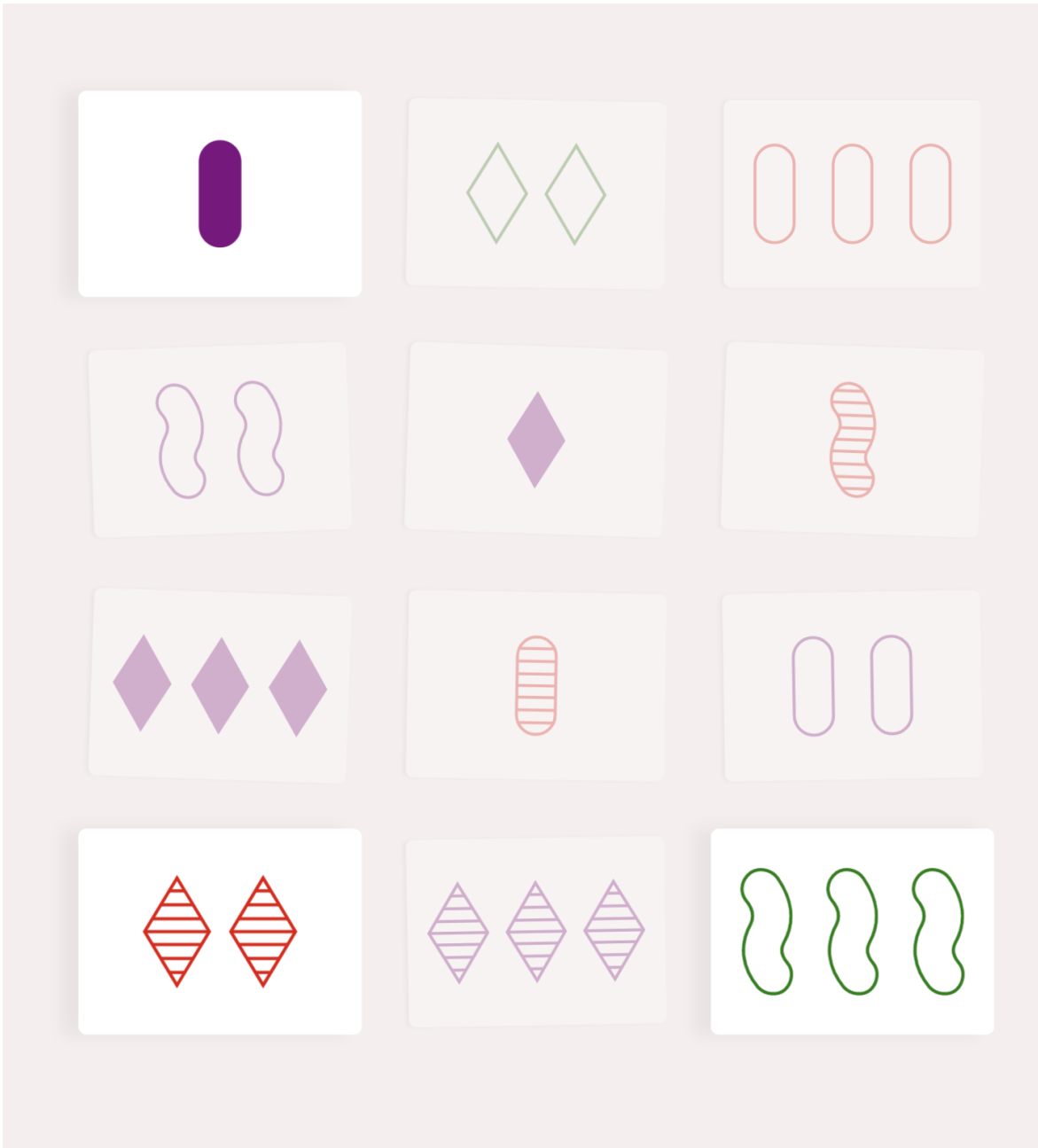
Objects $\mathcal{O}$ are Cartesian product of two sets of random attributes, strict ordering relations $a_1 \prec_a a_2 \prec_a a_3 \prec_a a_4$ and $b_1 \prec_b b_2 \prec_b \cdots \prec_b b_{12}$. Order relation uses $\prec_a$ as primary key and $\prec_b$ as secondary key; task is to sort 10 objects; non-overlapping testing dataset. Ablation model uses standard cross-attention Attention($Q \leftarrow S, K \leftarrow E, V \leftarrow E$).

Sample efficiency and abstraction

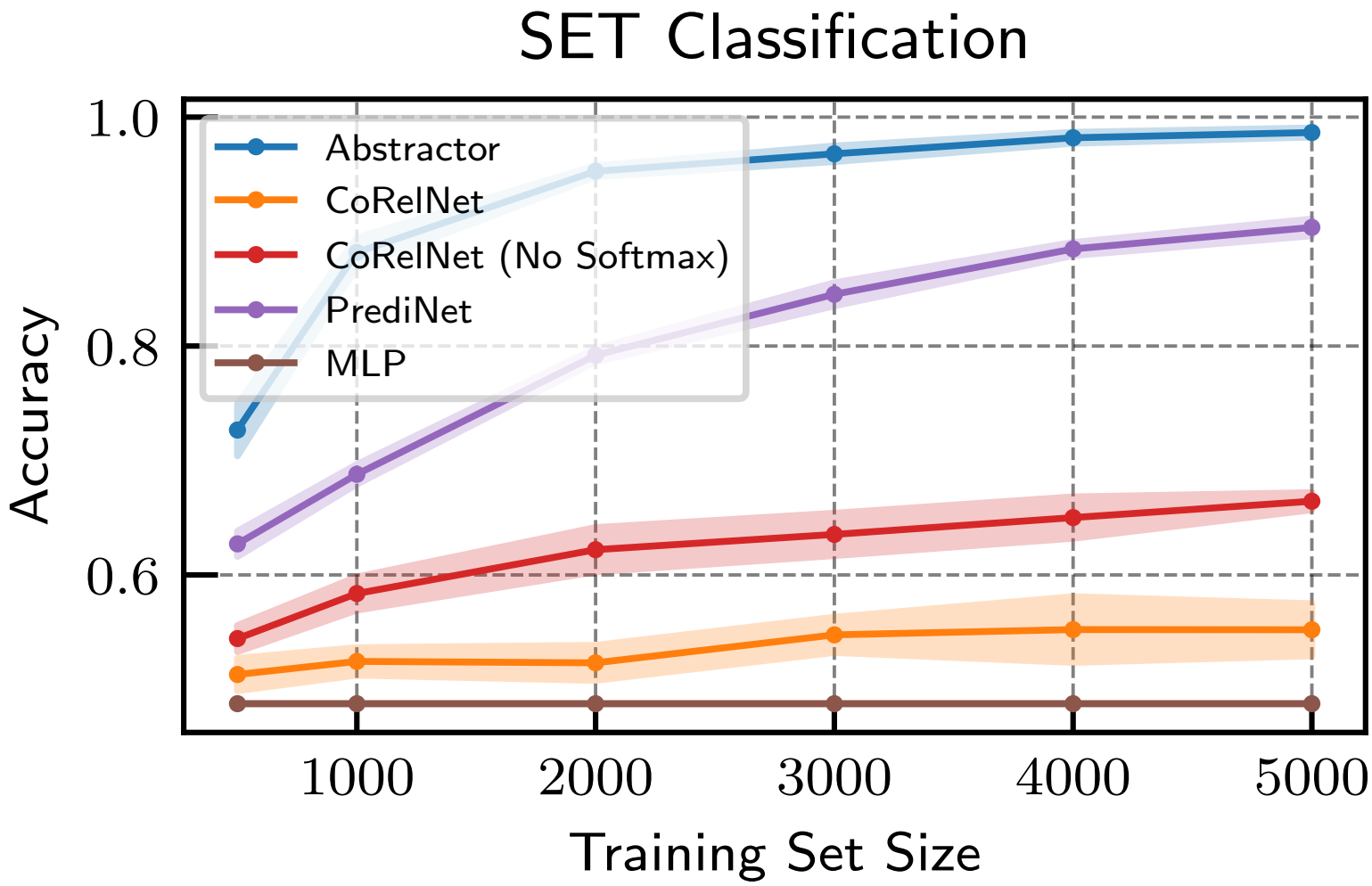


Objects $\mathcal{O}$ are Cartesian product of two sets of random attributes, strict ordering relations $a_1 \prec_a a_2 \prec_a a_3 \prec_a a_4$ and $b_1 \prec_b b_2 \prec_b \cdots \prec_b b_{12}$. Order relation uses $\prec_a$ as primary key and $\prec_b$ as secondary key; task is to sort 10 objects; non-overlapping testing dataset. Ablation model uses standard cross-attention Attention($Q \leftarrow S, K \leftarrow E, V \leftarrow E$).

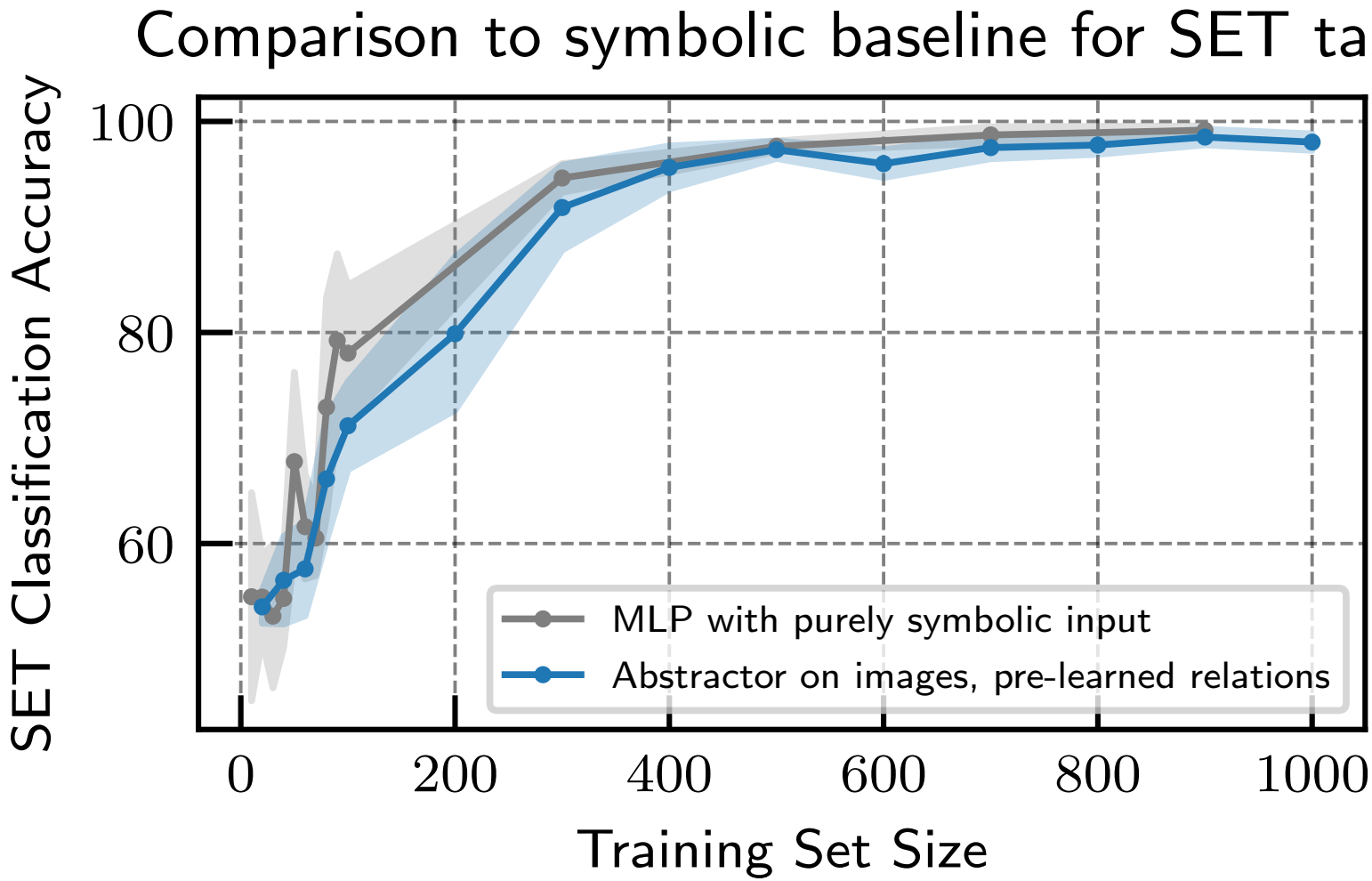




Abstractor on Set



Abstractor on Set

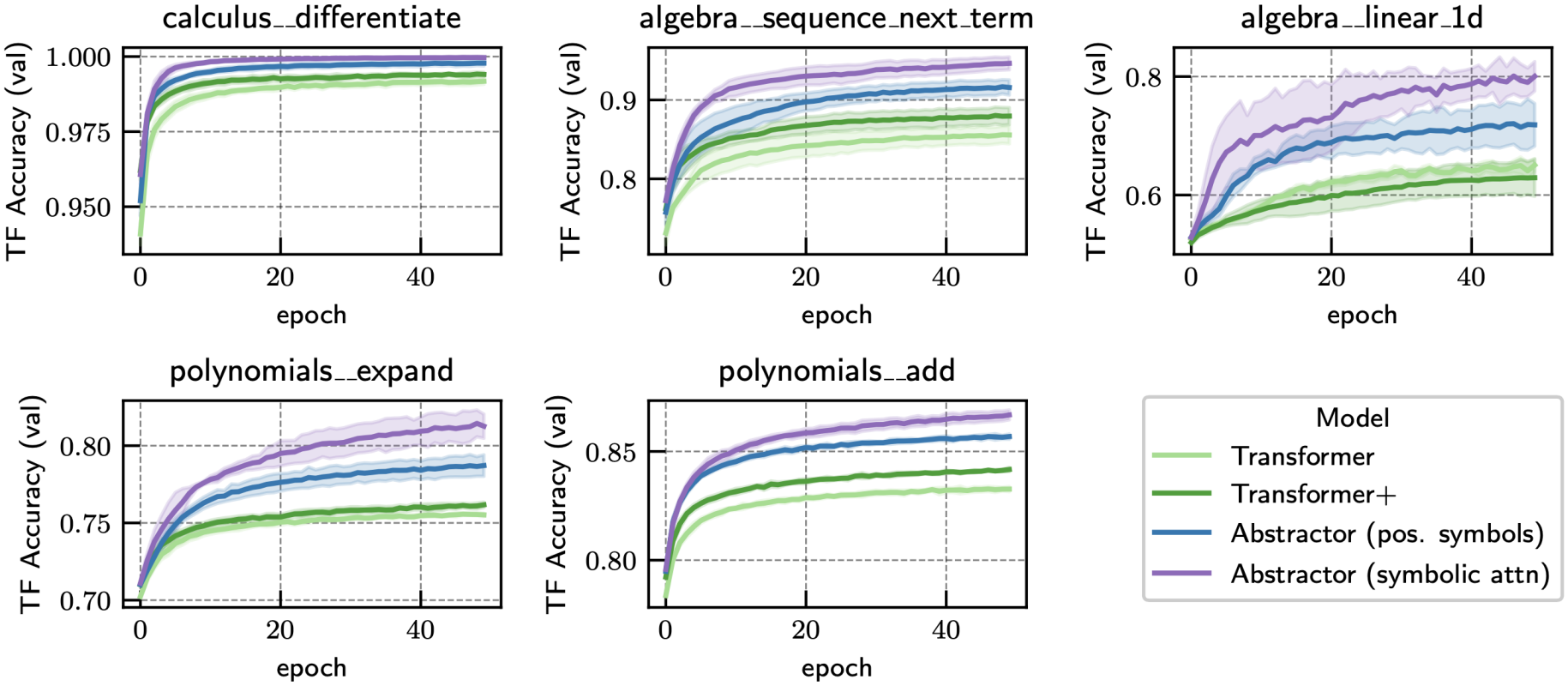


Trained on images of SET cards; CNN has two convolutional layers with 32 kernels of size 4 $\times$ 4 and max pooling; feature maps followed by MLP to embed to 32 dimensions. Abstractors trained separately to learn same/different relations for each attribute; used to initialize Abstractor with four relations to classify triples of cards. In symbolic model all attributes and relations between cards are hand coded in a binary feature vector.

Abstractor on math

Task: `polynomials__expand`
Question: `Expand (2*x + 3)*(x - 1).`
Answer: `2*x**2 + x - 3`

Task: `algebra__linear_1d`
Question: `Solve for z:  5*z + 2 = 9.`
Answer: `7/5`



Summary: Abstractors

- Emergent abstractions with distributed, neural representations
- Sample efficient: Learn much more quickly, with limited data
- Modular: Relations learned in one task can be used in another
- Computational models as hypotheses about the brain
- AI is likely to soon have advanced reasoning ability

arXiv:2304.00195