Disentangling and Integrating Relational and Sensory Information in Transformer Architectures

Awni Altabaa, John Lafferty

May 30, 2025

Yale University
arXiv:2405.16727, ICML '25

Big Picture: Why should we care about "relational reasoning"?

The Fundamental Principles of Intelligence

Hypothesis 0: Human & animal intelligence can be explained by a few core principles (rather than an encyclopedic list of heuristics)

Suggests the following goal: Study & uncover the inductive biases that humans & animals exploit to understand intelligence generally and inform design of AI

The Fundamental Principles of Intelligence

Deep learning systems themselves exploit several key inductive biases that underly their empirical success

Goal of AI Research: Uncover a core set of inductive biases for DL that enable data-efficient learning and reasoning over wide range of tasks and modalities

Hypothesis 1: Relational reasoning is one of these fundamental principles of intelligence

First: what *is* "relational reasoning"?

First: What is "relational reasoning"?

Reasoning about relationships between objects and how they interact in a given context/scene

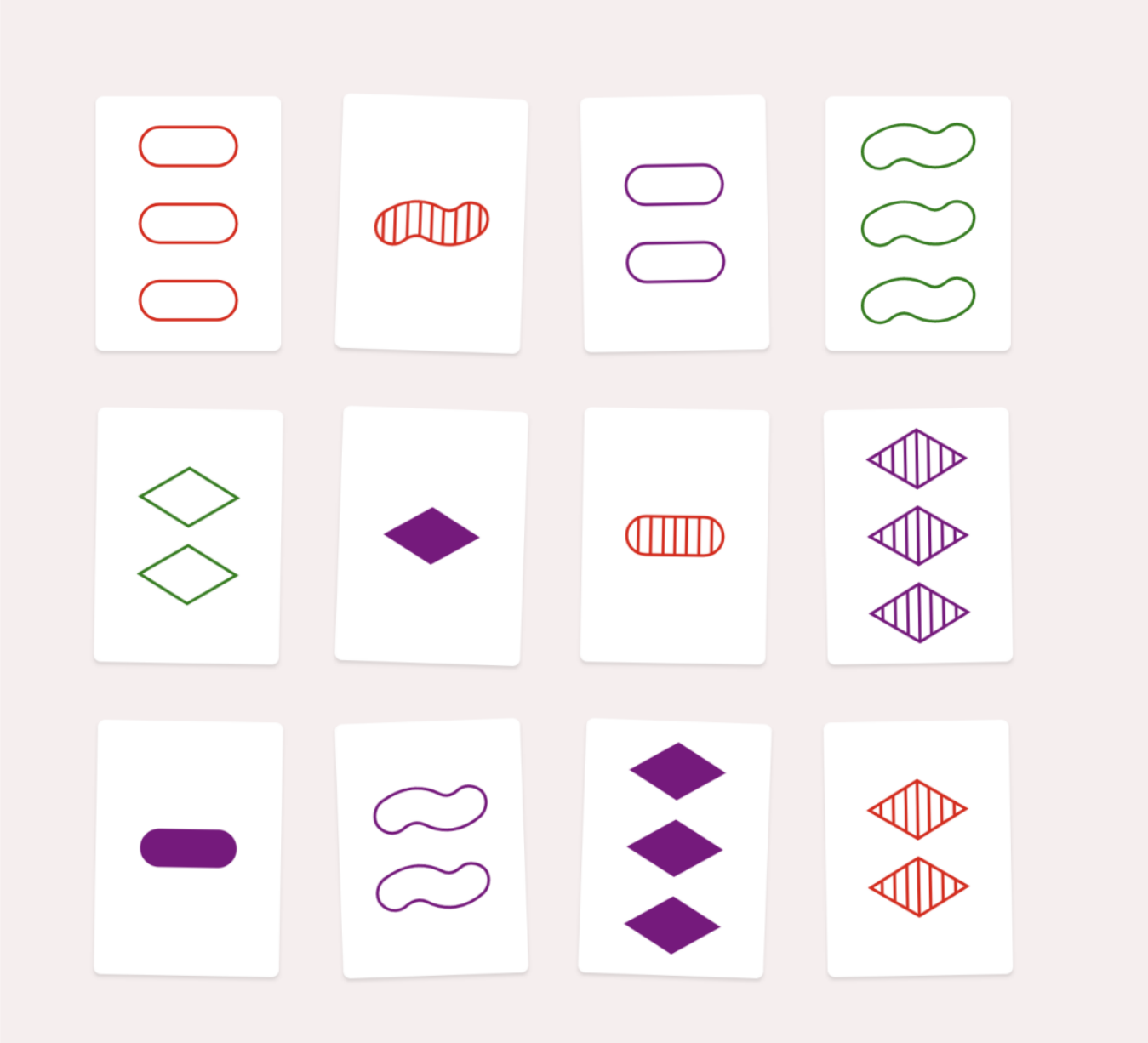
Perform comparisons under different attributes or features, at multiple levels of abstraction

Beyond recognizing individual objects by sensory pattern recognition; requires higher-order relationships

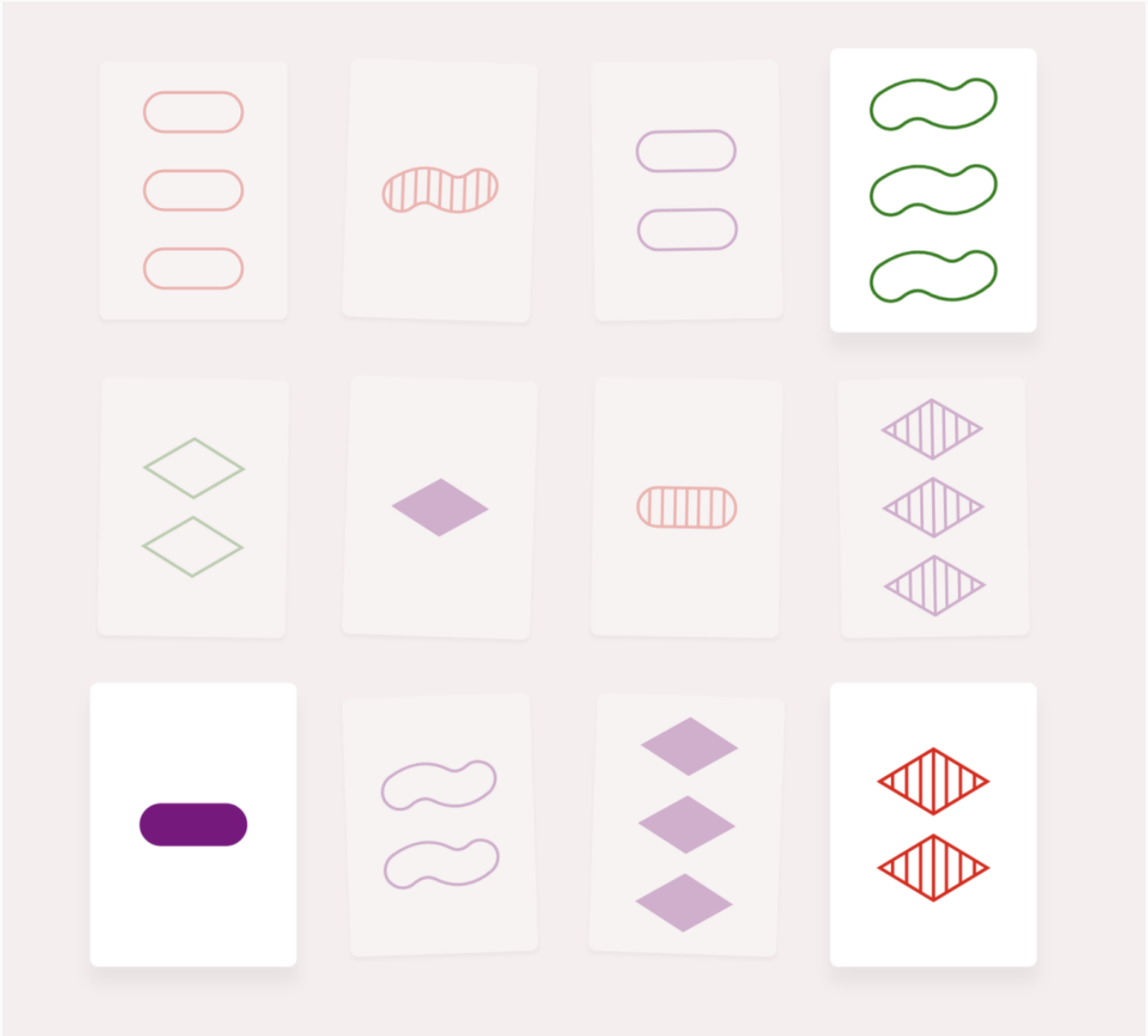
Clue to its importance: Humans have a natural ability (and a preference) to do relational reasoning

Let's walk through a couple simple illustrative examples of relational tasks

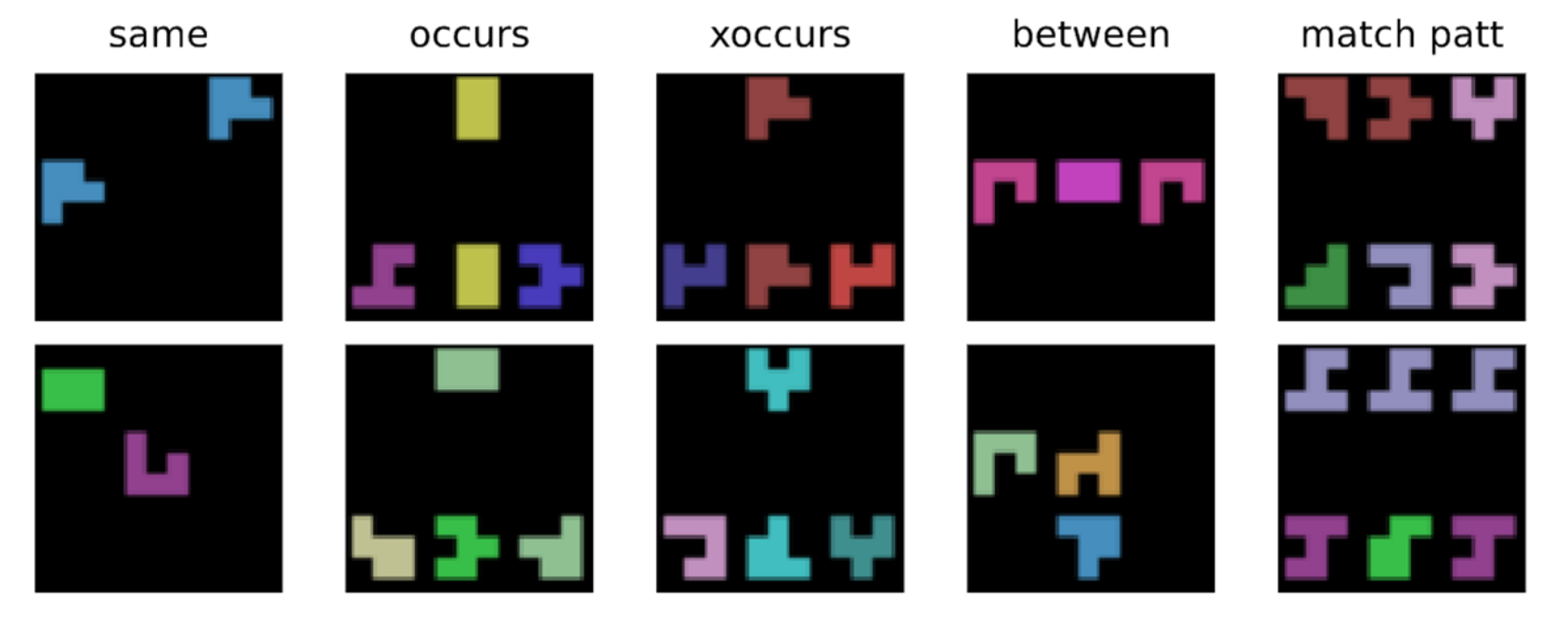
Example: *SET!* Card Game



Example: *SET!* Card Game



Example: Relational Games (Shanahan et al. 2020)



Relational Games tasks from Shanahan et al. (2020)

A Visual Relational Reasoning Task:
determine whether a particular relation holds or not

Returning to our original question:

Why should we care about relational reasoning?

Why should we care about "relational reasoning"?

A cornerstone of human intelligence

Underlies capabilities for

- *analogy*
- *abstraction*
- *generalization*

By relating new inputs to previously-seen stimuli, we form analogies and abstractions that allow us to systematically generalize.

"In the limit, relational reasoning yields universal inductive generalization from a finite and often very small set of observed cases to a potentially infinite set of novel instances ." — Goyal & Bengio (2022)

We'd like to take a step towards this central goal of AI research

Outline of Remainder of Talk

Big Picture: Why should we care about “relational reasoning”?

Main Idea & Goal

Transformers: The Sensory and the Relational

Relational Attention

Dual Attention Transformer Architecture

Empirical Investigation

Concluding Remarks

Main Idea & Goal

Our Goal

Our Goal: Make progress towards a universal neural architecture with explicit relational computational mechanisms & inductive biases

How to imbue Transformers with explicit relational inductive biases

- *Inductive Bias:* intrinsic preferences over solution space
- *View:* Two types
  - *Additive:* imbue architecture with mechanism, and let it learn to use it
  - *Subtractive:* constrain the space of representations a model can compute

Navigating the bitter lesson

- "*The Bitter Lesson*" — Rich Sutton
- Relational computational mechanisms parameterized by neural net & *learned*
  - scalable, general mechanisms;
  - avoid domain-specific heuristic, human-engineering
- The *versatility* of the Transformer architecture suggests it may form a powerful starting point

Some Lessons from Previous Work

Prior works on relational inductive biases

- Santoro et al. "*A simple neural network module for relational reasoning*' (2017)
- Shanahan et al. "*An Explicitly Relational Neural Network Architecture*" (2020)
- Kerg et al. "Inductive biases for relational tasks" (2022)
- Others...

Data-efficient relational reasoning requires inductive biases

- Standard neural models (e.g., Transformers) are *data-inefficient* at learning relational tasks; brittle OOD generalization
- Hypothesized Explanation: Neural networks overemphasize *individual object* representations while lacking explicit mechanisms for encoding and processing *relational* features.
- Common thread explored: constrain model to compute relational features—relational inductive biases

Tension: Generality vs. Inductive Biases

However, these models are narrow in domain

They improve relational processing, but lose generality

Empirical success limited to synthetic (purely relational) benchmarks

Our Goal: augment the Transformer architecture with explicit relational mechanisms & inductive biases

Transformers: The Sensory and the Relational

How to imbue Transformers with explicit relational inductive biases

Strength of Transformers: *attention*

Versatile information retrieval mechanism

Composable in *circuits* to carry out complex computation

(which we're now beginning to understand through systematic (mechanistic) interpretability work)

The Transformer architecture, essentially

Iterate two basic operations:

1. Information Retrieval: Attention

$$x_i' \leftarrow \sum_j \alpha_{ij}\, \phi_v(x_j)$$

2. Local Processing: Token-wise MLP

$$x_i' \leftarrow \mathrm{MLP}(x_i)$$

Attention, essentially

1. Compute attention scores

$$\alpha_{ij} = \mathrm{Softmax}([\langle \phi_q^{\mathrm{attn}}(x_i), \phi_k^{\mathrm{attn}}(x_j) \rangle]_{j=1}^{n})_j$$

2. Retrieve weighted combination of sensory values in context

$$\boldsymbol{e}_i \leftarrow \sum_j \alpha_{ij}\, \phi_v(x_j)$$

Two types of information

Fundamentally, attention is an information retrieval operation

Two key *types* of information

- Sensory: features or attributes of individual objects
- Relational: relationships between objects

Standard attention captures the former, but not the latter

Two types of attention

Correspondingly, there ought to be two types of attention

(Standard) Sensory Attention:

retrieval of *sensory* information in context

Relational Attention:

retrieval of *relational* information in context

Relational Attention

High-level: Relational Attention

1. Attend
2. Relate
3. Tag with symbols

1) Attention

Same as standard (sensory attention)

Compute attention scores via learned query/key maps

$$\alpha_{ij} = \mathrm{Softmax}([\langle \phi_q^{\mathrm{attn}}(x_i), \phi_k^{\mathrm{attn}}(x_j) \rangle]_{j=1}^{n})_j$$

2) Computing Relations

Relation vector, representing a series of comparisons under different attributes or extracted features

Computed as a series inner products under different learned feature maps

$$\boldsymbol{r}_{ij} = \left(\left\langle \phi^{\text{rel}}_{q,\ell}(x_i), \phi^{\text{rel}}_{k,\ell}(x_j) \right\rangle \right)_{\ell \in [d_r]} \in \mathbb{R}^{d_r}$$

3) Symbols

Tag each object in the context with an *symbol*

$$(s_1, \ldots, s_n) = \text{SymbolRetriever}(x_1, \ldots, x_n)$$

Serve as reference/pointer/identifier of selected object with whom the relation is with, abstracted away from high-dimensional sensory features

We experiment with different symbol assignment mechanisms: positional, relative positional, “soft-equivalence class”

Relational Attention: Putting it all together

Putting it all together

$$\boldsymbol{a}_i \leftarrow \sum_j \alpha_{ij} \cdot (W_r \, \boldsymbol{r}_{ij} + W_s \, \boldsymbol{s}_j)$$

α_{ij} : attention scores — govern selection criterion

$\boldsymbol{r}_{ij}$: relation vector — relational information

$\boldsymbol{s}_j$: symbol — identifier of source/sender object

W_r, W_s : learned linear maps — organize information in residual stream

A few comments...

Causal masking

Positional encoding

Symmetric relations

Computational complexity

Dual Attention Transformer Architecture

Dual Attention Transformer

Relational attention : a mechanism for routing relational information

Both *sensory* and *relational* information are crucial for reasoning over collections or sequences of objects.

Dual Attention Transformer (DAT): A variant of the Transformer architecture that routes both types of information in the information retrieval step.

Introduces explicit relational processing mechanisms, while retaining sensory processing capabilities.

Dual Attention

Dual Attention is a variant of multi-head attention with *two types of attention heads*: *sensory* and *relational* .

Dual Attention

Algorithm 1: Dual Attention

Input: $\boldsymbol{x} = (x_1, \ldots, x_n) \in \mathbb{R}^{n \times d}$

Compute self-attention heads

$$\boldsymbol{\alpha}^{(h)} \leftarrow \mathrm{Softmax}\big((\boldsymbol{x}\, W^{\mathrm{attn}}_{q,h})(\boldsymbol{x}\, W^{\mathrm{attn}}_{k,h})^{\intercal}\big), \quad h \in [n^{sa}_h]$$

$$e^{(h)}_i \leftarrow \sum_j \alpha^{(h)}_{ij} x_j\, W^h_v, \quad i \in [n], h \in [n^{sa}_h]$$

$$\boldsymbol{e}_i \leftarrow \mathrm{concat}\big(e^{(1)}_i, \ldots, e^{(n^{sa}_h)}_i\big)\, W^{sa}_o, \quad i \in [n]$$

Assign symbols:

$$\boldsymbol{s} = (s_1, \ldots, s_n) \leftarrow \mathrm{SymbolRetriever}(\boldsymbol{x};\, S_{\mathrm{lib}})$$

Compute relational attention heads

$$\boldsymbol{\alpha}^{(h)} \leftarrow \mathrm{Softmax}\big((\boldsymbol{x}\, W^{\mathrm{attn}}_{q,h})(\boldsymbol{x}\, W^{\mathrm{attn}}_{k,h})^{\intercal}\big), \quad h \in [n^{ra}_h]$$

$$\boldsymbol{r}_{ij} \leftarrow \big(\langle x_i\, W^{\mathrm{rel}}_{q,\ell}, x_j\, W^{\mathrm{rel}}_{k,\ell}\rangle\big)_{\ell \in [d_r]} \quad i, j \in [n]$$

$$\boldsymbol{a}^{(h)}_i \leftarrow \sum_j \alpha^{(h)}_{ij} \big(\boldsymbol{r}_{ij}\, W^h_r + s_j\, W^h_s\big), \quad i \in [n],\ h \in [n^{ra}_h]$$

$$\boldsymbol{a}_i \leftarrow \mathrm{concat}\big(a^{(1)}_i, \ldots, a^{(n^{ra}_h)}_i\big)\, W^{ra}_o, \quad i \in [n]$$

Output: $\big(\mathrm{concat}(\boldsymbol{e}_i, \boldsymbol{a}_i)\big)_{i=1}^{n}$

Dual Attention Transformer: Encoder & Decoder

Algorithm 2: Dual Attention Encoder Block

Input: $\boldsymbol{x} \in \mathbb{R}^{n \times d}$

$\boldsymbol{x} \leftarrow \mathrm{Norm}(\boldsymbol{x} + \mathrm{DualAttn}(\boldsymbol{x}))$

$\boldsymbol{x} \leftarrow \mathrm{Norm}(\boldsymbol{x} + \mathrm{MLP}(\boldsymbol{x}))$

Output: $\boldsymbol{x}$

Algorithm 3: Dual Attention Decoder Block

Input: $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^{n \times d}$

$\boldsymbol{x} \leftarrow \mathrm{Norm}(\boldsymbol{x} + \mathrm{DualAttn}(\boldsymbol{x}))$

$\boldsymbol{x} \leftarrow \mathrm{Norm}(\boldsymbol{x} + \mathrm{CrossAttn}(\boldsymbol{x}, \boldsymbol{y}))$

$\boldsymbol{x} \leftarrow \mathrm{Norm}(\boldsymbol{x} + \mathrm{MLP}(\boldsymbol{x}))$

Output: $\boldsymbol{x}$

Empirical Investigation

Prelude: what questions are we trying to answer?

How does the *DAT* perform on synthetic relational benchmarks?

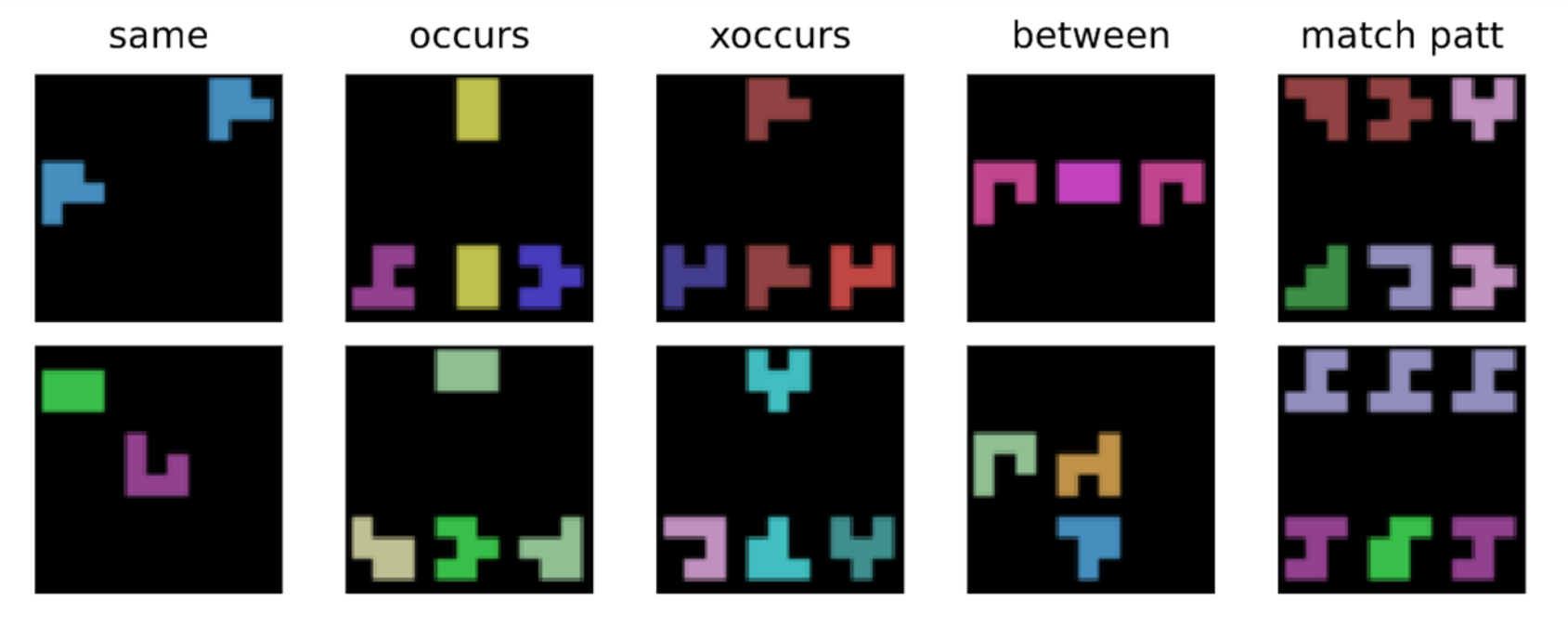
Data efficiency

Scalability with data and model size (recall: bitter lesson)

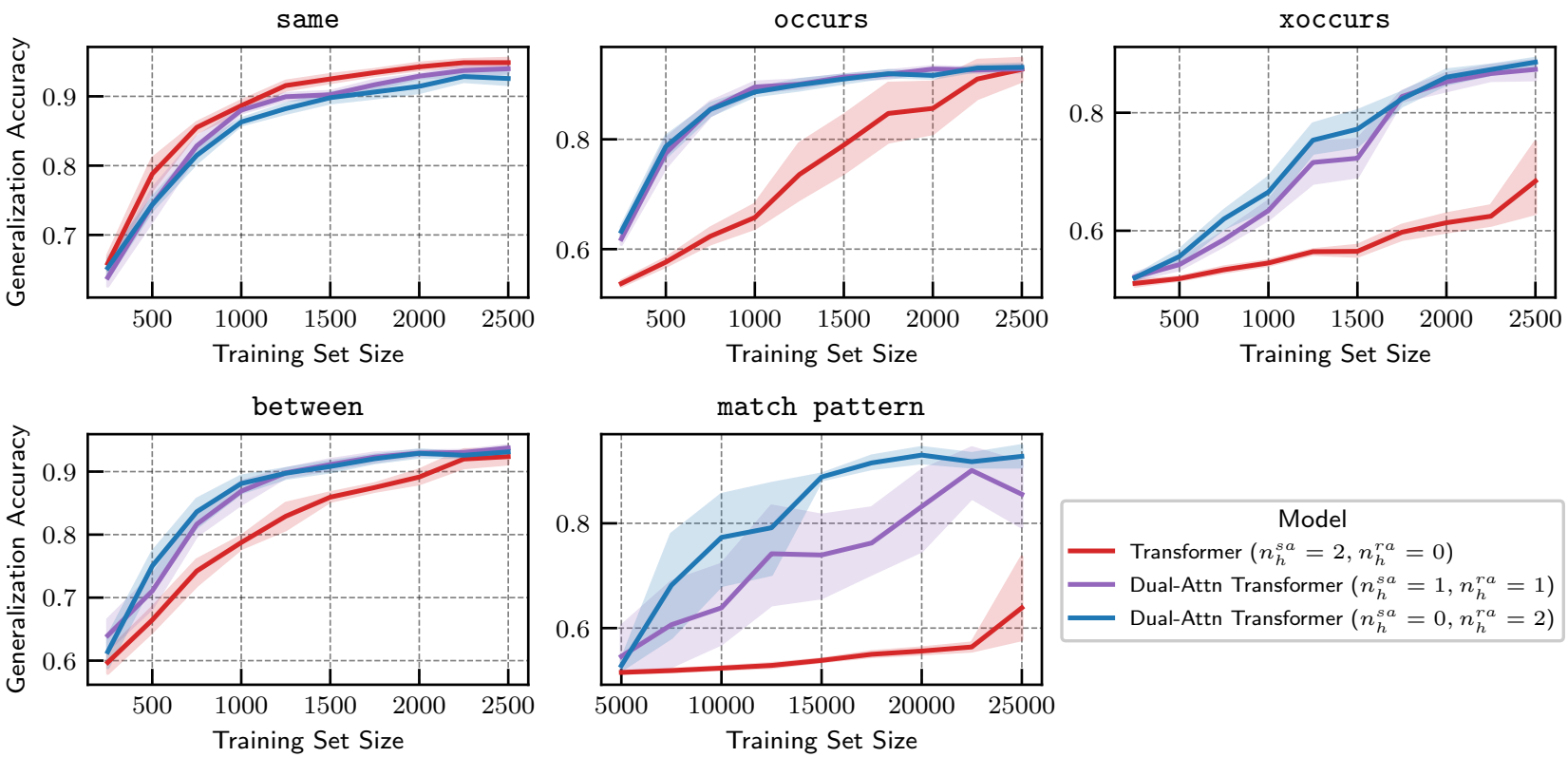
Applicability to complex real-world tasks; versatility across data modalities (language & vision)

Synthetic Relational Benchmarks: Relational Games (Shanahan et al. 2020)

Synthetic Relational Tasks: Task



Synthetic Relational Tasks: Results



Mathematical Problem-Solving (Seq2Seq)

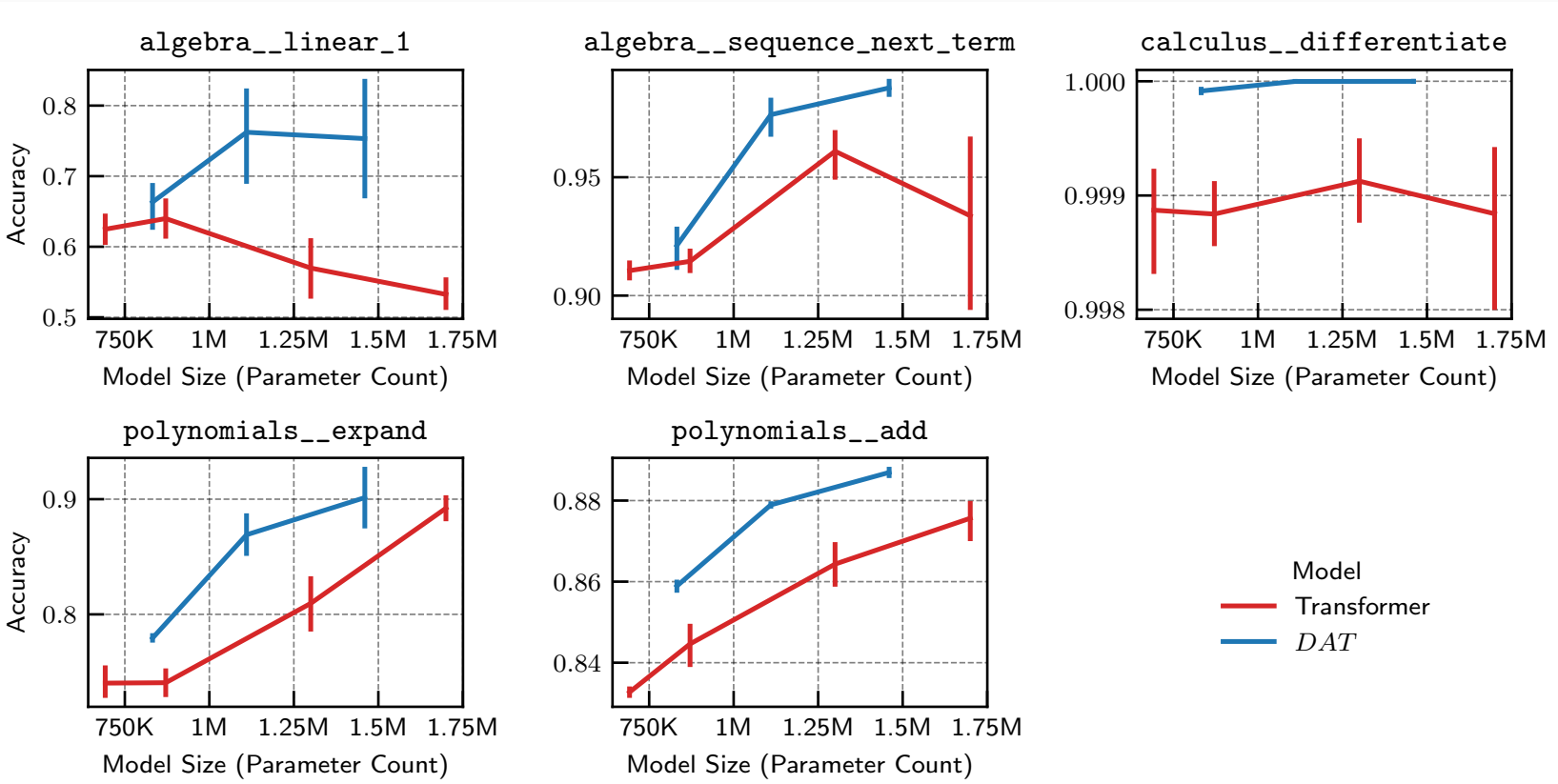
Mathematical Problem-Solving (Seq2Seq): Task

Dataset due to Saxton et al. (2019)

Modeled as char-level Sequence-to-Sequence task with *encoder-decoder* architecture

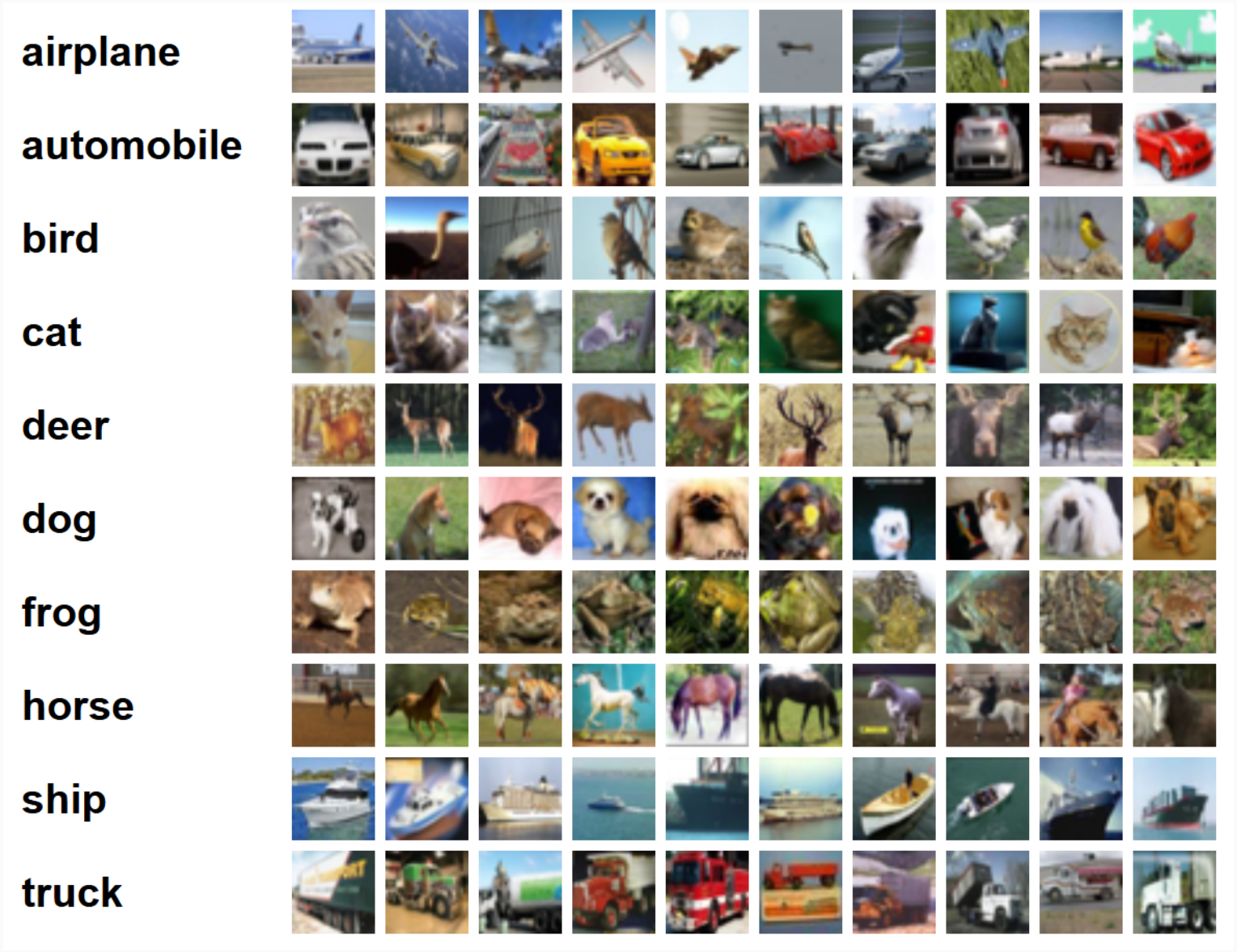
Module	Math Dataset Example
algebra_linear_1d	**Q:** Solve for x: $3x + 7 = 19$ **A:** $x = 4$
algebra_sequence_next_term	**Q:** What is the next term in the sequence $2, 5, 8, 11, \ldots$? **A:** 14
calculus_differentiate	**Q:** Find the derivative of $f(x) = 3x^2 + 2x - 5$ with respect to x. **A:** $6x + 2$
polynomials_expand	**Q:** Expand $(2x + 3)(x - 1)$. **A:** $2x^2 + x - 3$
polynomials_add	**Q:** Add the polynomials: $(2x^2 + 3x + 1) + (x^2 - 2x + 4)$ **A:** $3x^2 + x + 5$

Mathematical Problem-Solving (Seq2Seq): Results



Visual Processing (CIFAR)

Visual Processing (CIFAR): Task



Visual Processing (CIFAR): Results

ViT-style encoder-only architecture processing image as sequence of patches

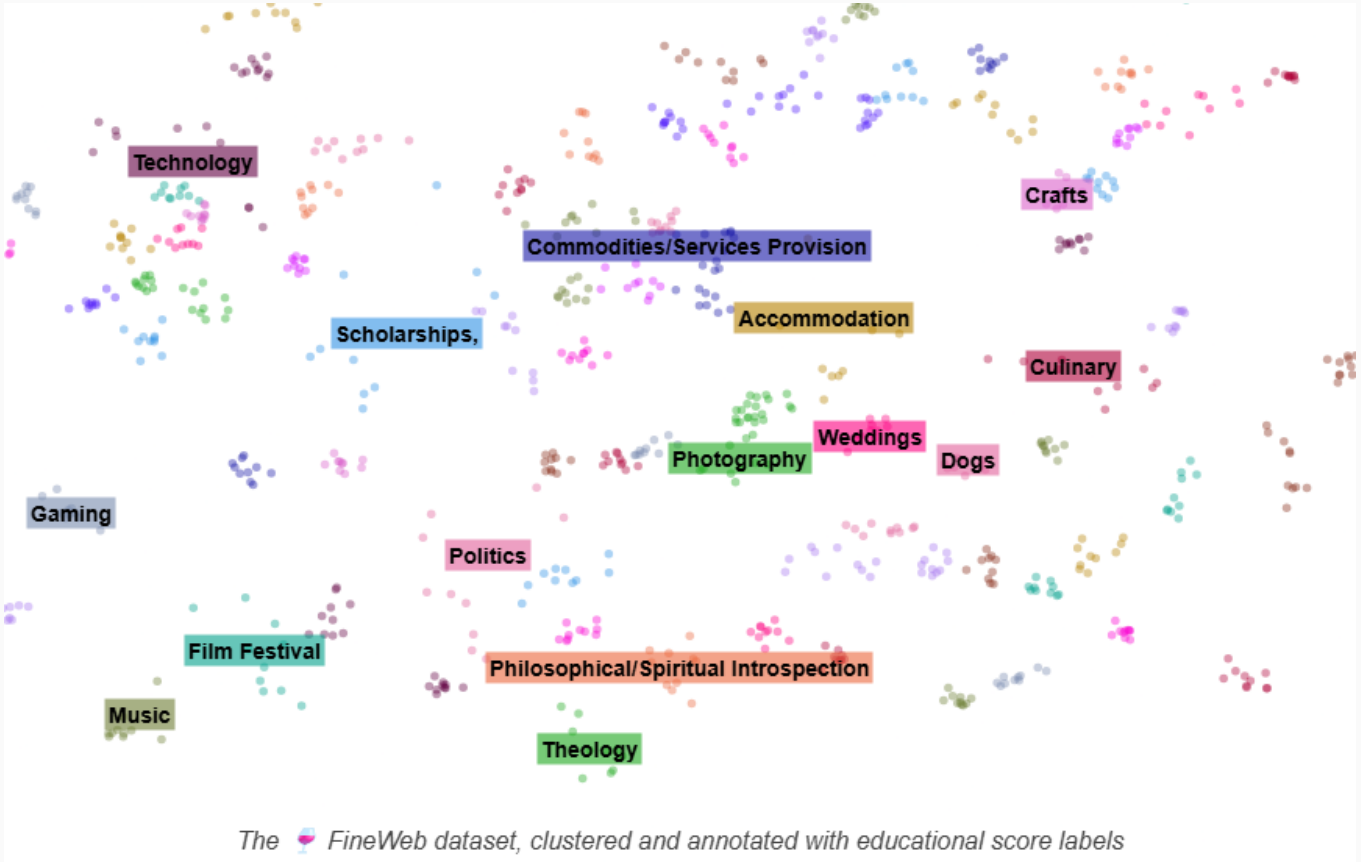
Dataset	Model	Params	Accuracy
CIFAR-10	*ViT*	7.1M	$86.4 \pm 0.1\%$
	ViDAT	6.0M	$\mathbf{89.7 \pm 0.1\%}$
CIFAR-100	*ViT*	7.2M	$68.8 \pm 0.2\%$
	ViDAT	6.1M	$\mathbf{70.5 \pm 0.1\%}$

Language Modeling

Language Modeling: Task

Autoregressive causal language modeling with a "decoder-only" architecture

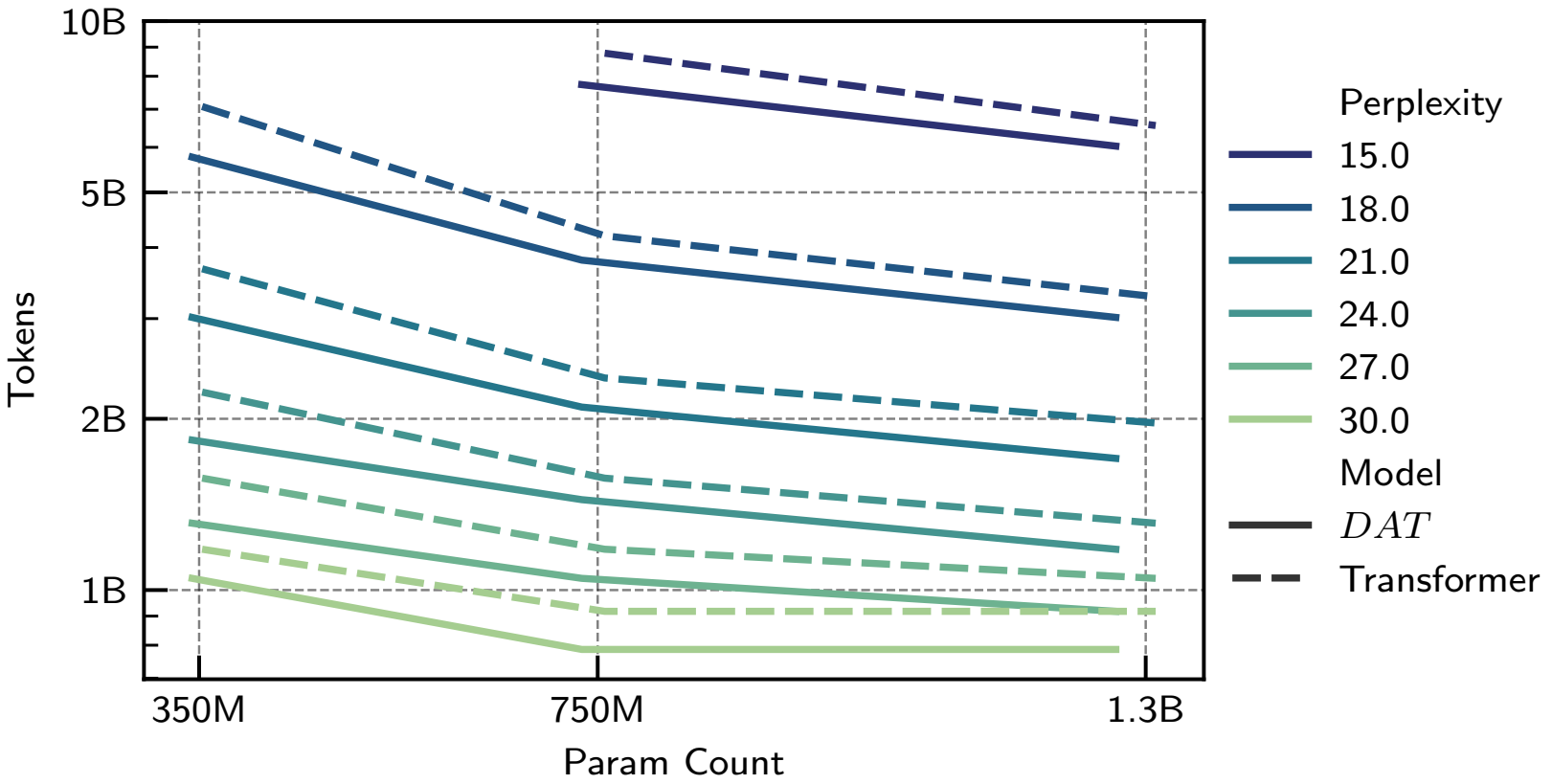
Use the Fineweb-Edu dataset (curated high-quality text data); train on 10B tokens



The FineWeb dataset, clustered and annotated with educational score labels

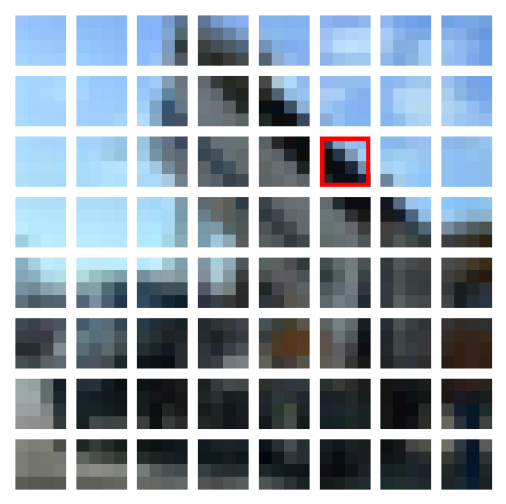
Language Modeling: Results

Evaluate scaling with data and model size



A bit of visualization/interpretation

Interpreting *ViDAT* model

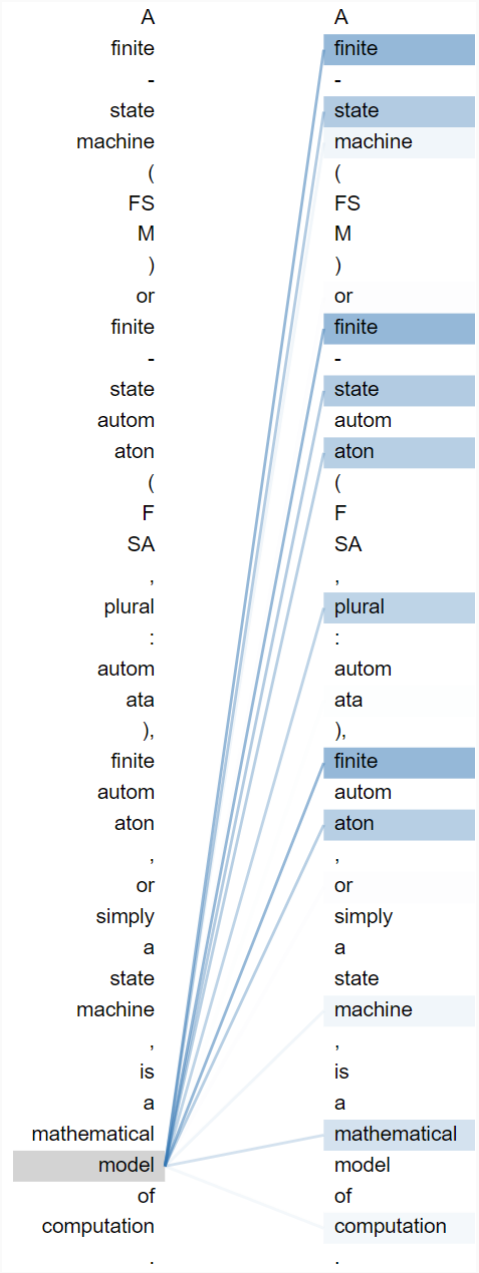


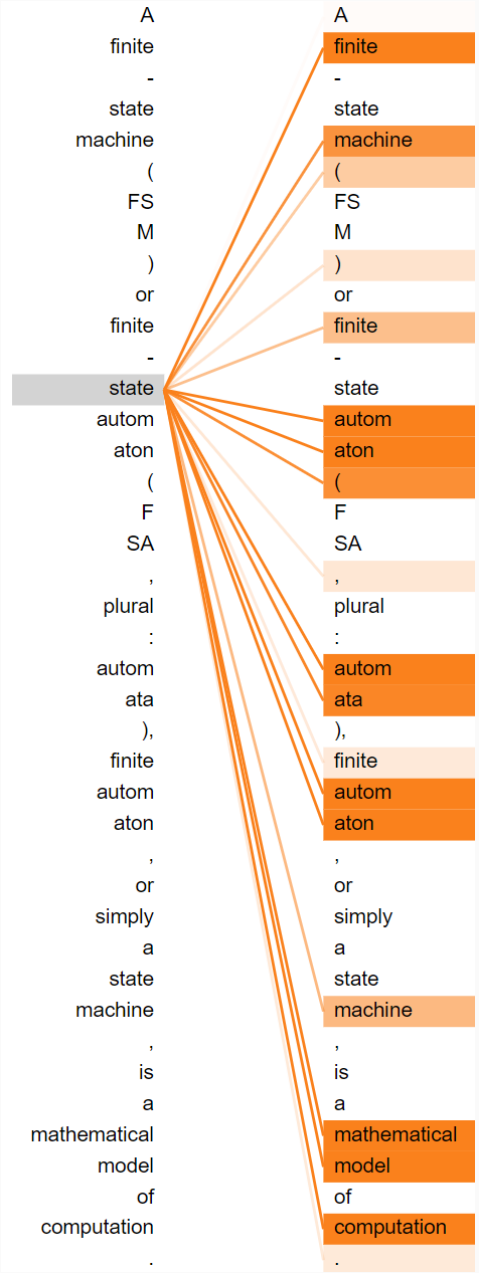
(a) Original Image

(b) A Relation in the First Layer

(c) A Relation in the Fifth Layer

Interpreting *DAT* Language Models





Concluding Remarks

Concluding Remarks

Relational reasoning is a core facet of human intelligence, underpinning abilities for analogy, abstraction, and generalization

It is likely an important component of artificial intelligence as well

In this work, we took a step towards developing neural architectures with enhanced relational processing capabilities, while retaining powerful sensory processing

Concluding Remarks: Future Work

Interpretability:

- How is *DAT* learning to use its relational processing mechanisms?
- Can specific “circuits” be identified?
- How does *DAT* achieve improved data efficiency in different tasks?

Iterate & tweak architecture; find good choices for hyperparameters

Computational considerations: optimize implementation

Thank You

Discussion Time...

- Joint work with *John Lafferty*
- Supported by funding from ARNI NSF AI Institute
- Paper: *arXiv:2405.16727* / ICML '25
- Project webpage: *https://awni.xyz/dual-attention/*
  - Open weights on HF (DAT-LM up to 1.3B-params)
  - Implementation available via python package
    pip install dual-attention
- Personal webpage: *https://awni.xyz*