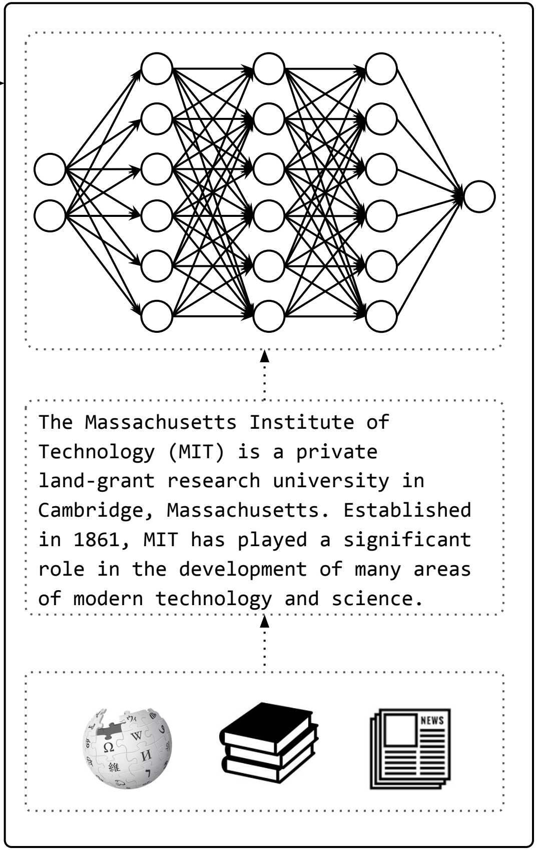
Combining Symbolic Logic with Transformers for Interpretable and Controllable LLMs

The “LLM Paradigm”

Text

Language Modeling

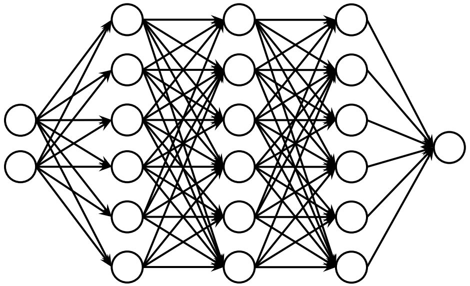


Language model pretraining for learning linguistic/world knowledge.

The “LLM Paradigm”

Text → Language Modeling → Applications

Language Modeling



The Massachusetts Institute of Technology (MIT) is a private land-grant research university in Cambridge, Massachusetts. Established in 1861, MIT has played a significant role in the development of many areas of modern technology and science.

Applications

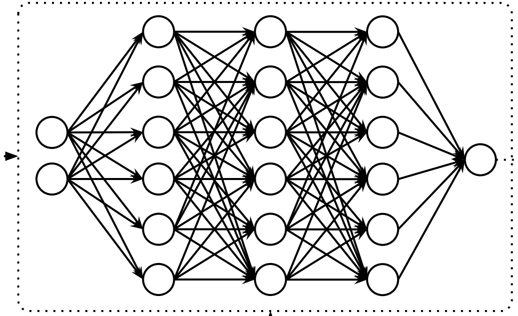
- Machine Translation
- Question Answering
- Semantic Parsing
- Dialogue
- ⋮

Language model pretraining for learning linguistic/world knowledge.

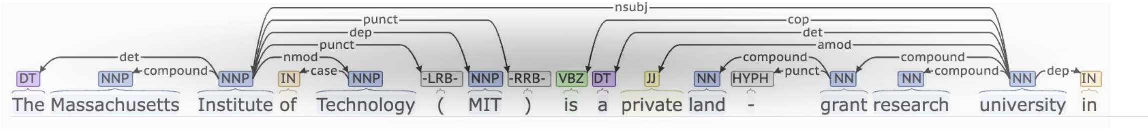
Focus on adapting pretrained language models (finetuning, prompting, etc.).

LLMs & Implicit Structures

```
The Massachusetts Institute
of Technology (MIT) is a
private land-grant research
university in ???
```



```
Cambridge
```



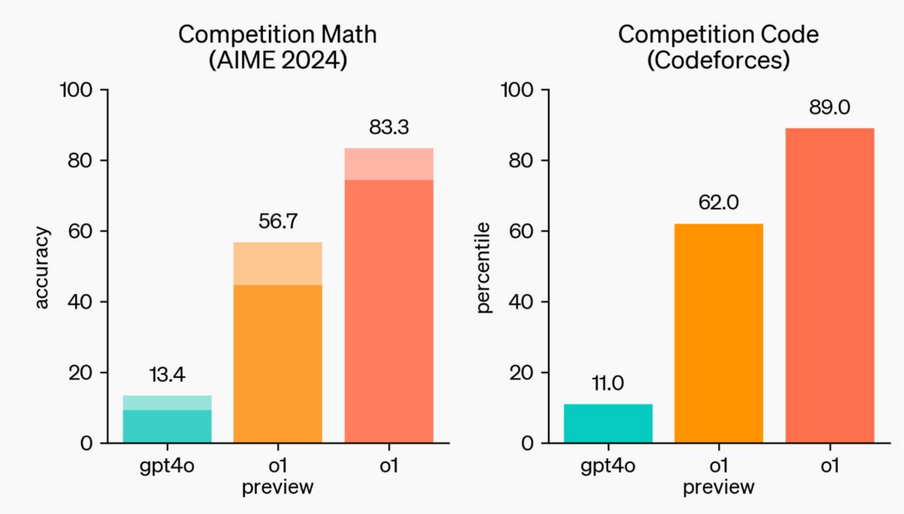
The language modeling (next-word prediction) objective forces the LLM to *implicitly* encode useful structure.

LLMs & *Explicit* Structures?

Is language modeling "all you need"?

$$\arg\max_{\theta} \; P_{\theta}(\boldsymbol{x}_t \mid \boldsymbol{x}_{<t})$$

Maybe…



LLMs & *Explicit* Structures?

Alice and Bob play the following game. A stack of n tokens lies before them. The players take turns with Alice going first. On each turn, the player removes either 1 token or 4 tokens from the stack. Whoever removes the last token wins. Find the number of positive integers n less than or equal to 2024 for which there exists a strategy for Bob that guarantees that Bob will win the game regardless of Alice's play.

The problem asks for the number of positive integers $n \leq 2024$ such that Bob has a winning strategy. This is a typical impartial game that can be analyzed using the Sprague-Grundy theorem or by directly finding the winning and losing positions. Let W be the set of winning positions (number of tokens n such that the player whose turn it is can force a win) and L be the set of losing positions. [...]

A quantum mechanical particle of mass m moves in two dimensions in the following potential, as a function of the polar coordinates (r, θ):

$$V(r,\theta) = \frac{1}{2}kr^2 + \frac{3}{2}kr^2\cos^2(\theta)$$

Find the energy spectrum. Hint: Write the potential in Cartesian coordinates.

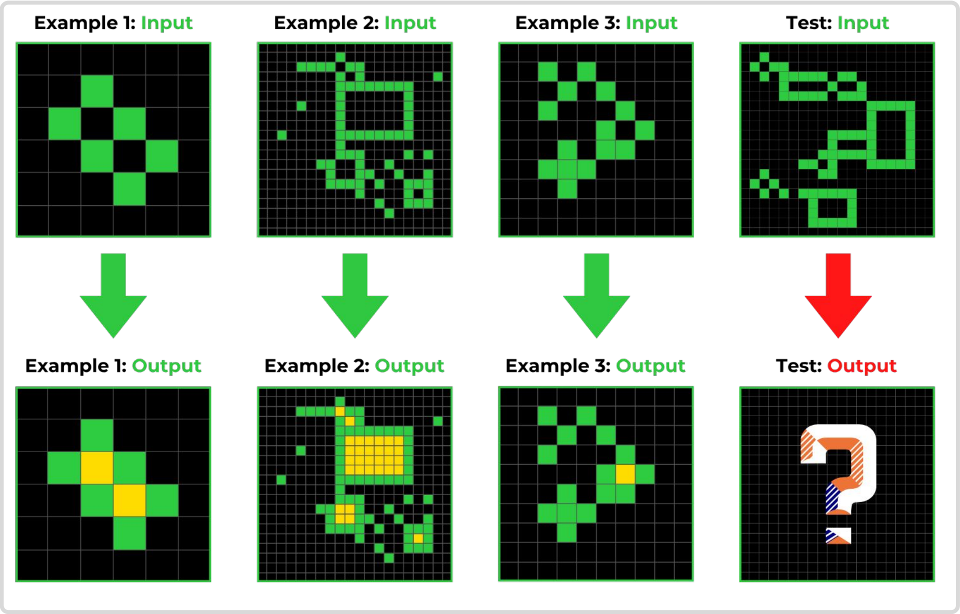
The problem asks for the energy spectrum of a particle in a given 2D potential. The potential is given in polar coordinates (r, θ):

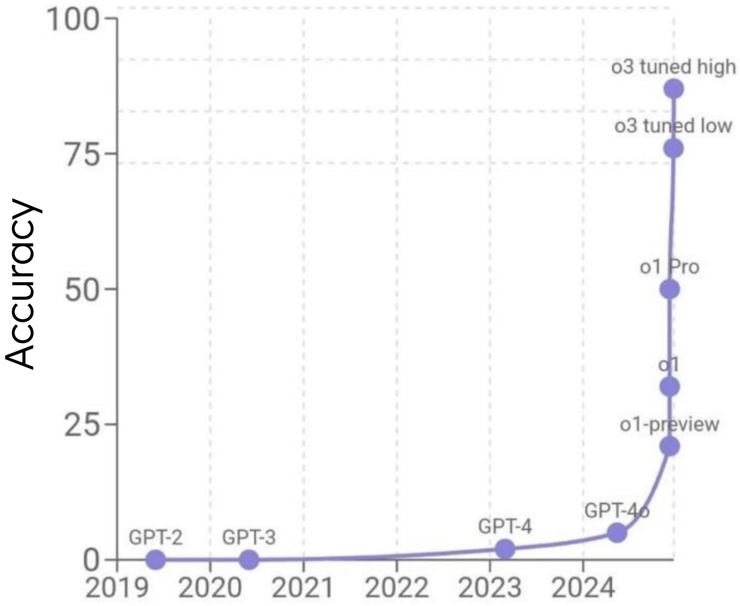
49. $$V(r,\theta) = \frac{1}{2}kr^2 + \frac{3}{2}kr^2\cos^2(\theta).$$

We can rewrite the potential as:

LLMs & *Explicit* Structures?

OpenAI o3 (Dec 2024) on ARC Challenge





LLMs & *Explicit* Structures?

OpenAI o3 (Dec 2024) on ARC Challenge

Nature

How should we test AI for human-level intelligence? OpenAI's o3 electrifies quest

The technology firm OpenAI made headlines last month when its latest experimental chatbot model, o3, achieved a high score on a test that...

MSN

OpenAI's o3 system has reached human level on a test for 'general intelligence'

OpenAI's o3 system got a score of 85% from the previous 55% on the ARC-AGI benchmark. This score is on par with the average human score.

YO

Can you generate a sentence with 5 occurrences of the word "the"?

Certainly! Here's a sentence with 5 occurrences of the word "the":

"The cat, which the little girl found in the alley, was the fluffiest of the strays."

YO

What about 15 occurrences?

Of course! Here's a sentence with 15 occurrences of the word "the":

"The baker, whom the townspeople considered the finest in the region, believed that the flour from the mill by the river, which the locals called the lifeblood of the valley, was the secret behind the exquisite texture of the bread that the citizens, the tourists, and the critics all loved the most."

YO

What about 15 occurrences?

Of course! Here's a sentence with 15 occurrences of the word "the":

"The baker, whom the townspeople considered the finest in the region, believed that the flour from the mill by the river, which the locals called the lifeblood of the valley, was the secret behind the exquisite texture of the bread that the citizens, the tourists, and the critics all loved the most."

YO

I count 17 occurrences :(

Apologies for the oversight. Here's a revised sentence with exactly 15 occurrences of the word "the":

"In the heart of the village, the market buzzed with the chatter of the locals, while the children played near the fountain, the tourists browsed through the crafts, and the elders told stories from the past, creating the tapestry of the community's life."

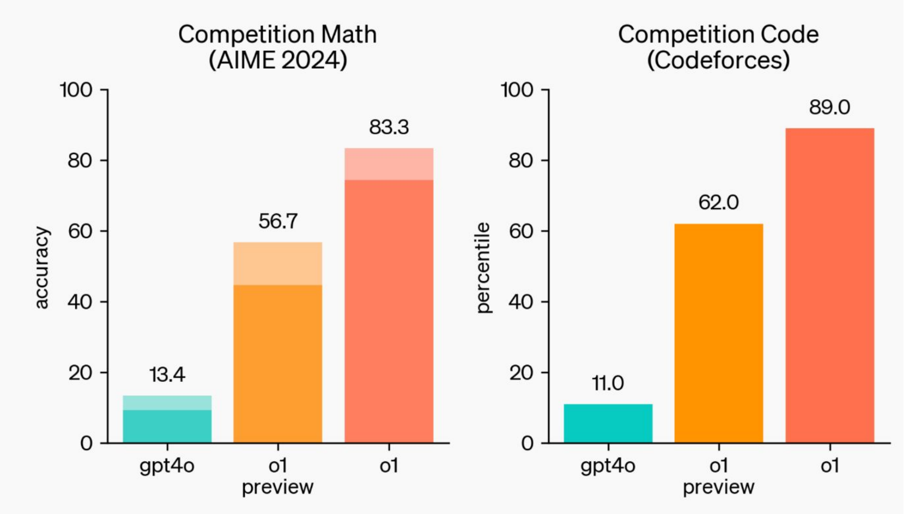
LLMs & *Explicit* Structures?

Is language modeling "all you need"?

$$\arg\max_{\theta} \; P_\theta(\boldsymbol{x}_t \,|\, \boldsymbol{x}_{<t})$$

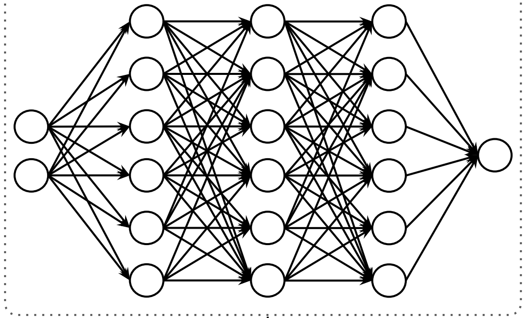
Maybe…

But not quite yet!



LLMs vs. Symbolic Systems

LLMs

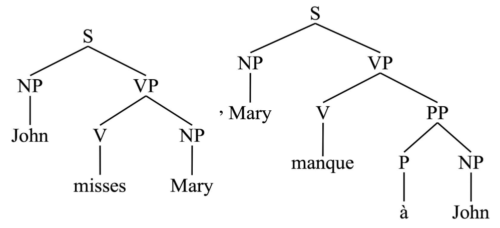


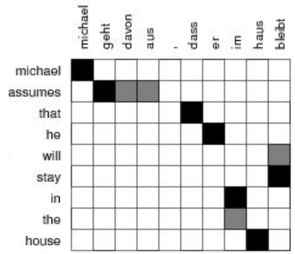
```
English: Hello, how are you?
French: Bonjour, comment ça va?
Spanish: Hola, ¿cómo estás?
German: Hallo, wie geht es dir?
Italian: Ciao, come stai?
```

✔ Generality/Flexibility ✗

✗ Interpretability/Controllability ✔

Symbolic Systems





LLMs + Symbolic Systems?

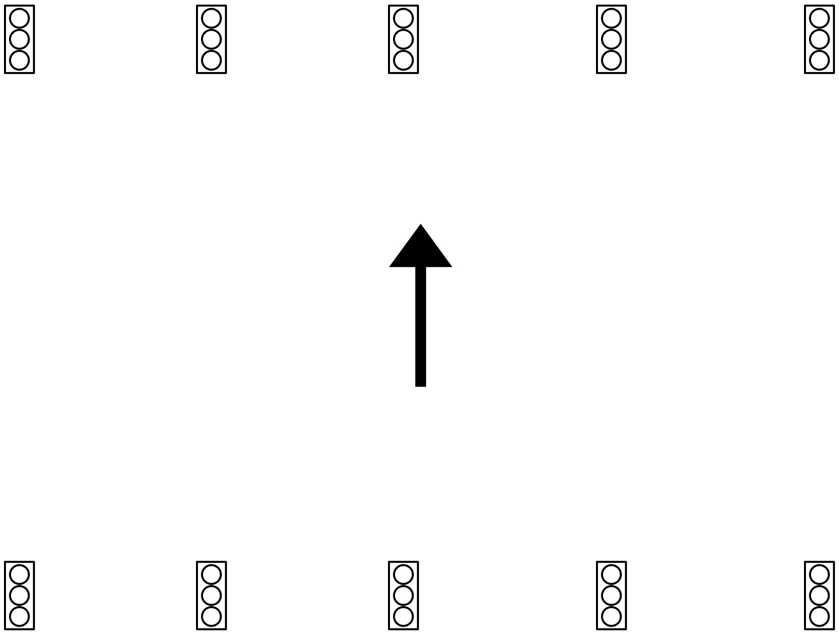
How can we combine the controllability/interpretability of symbolic models with the flexibility of LLMs?

How do Transformers work?

L : sequence length

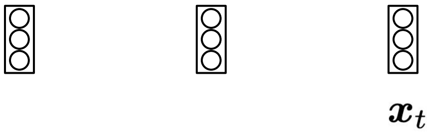
d : hidden state dimension

Output vectors $\mathbf{O} \in \mathbb{R}^{L \times d}$



Input vectors $\mathbf{X} \in \mathbb{R}^{L \times d}$

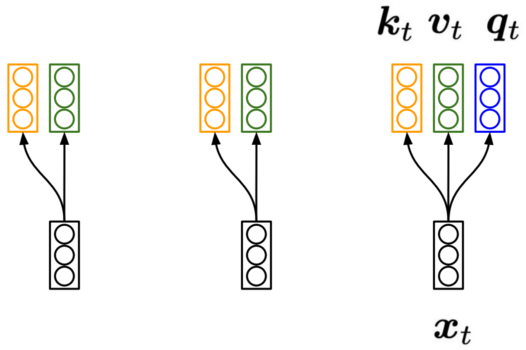
How do Transformers work?



How do Transformers work?

$$\boldsymbol{q}_t,\ \boldsymbol{k}_t,\ \boldsymbol{v}_t = \boldsymbol{x}_t\boldsymbol{W}_Q,\ \boldsymbol{x}_t\boldsymbol{W}_K,\ \boldsymbol{x}_t\boldsymbol{W}_V$$

Key K
Value V
Query Q



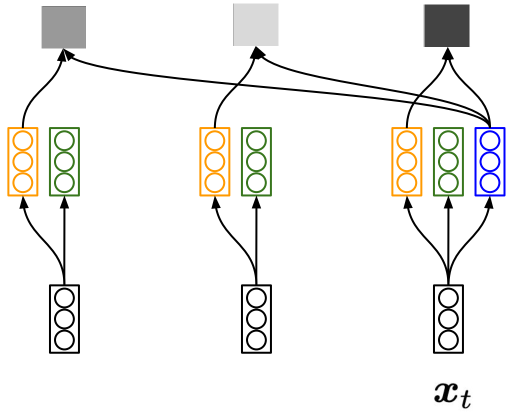
How do Transformers work?

“Attention distribution”: Model pairwise interactions between current and previous inputs

$$\frac{\exp(\boldsymbol{q}_t^\top \boldsymbol{k}_j)}{\sum_{l=1}^{t} \exp(\boldsymbol{q}_t^\top \boldsymbol{k}_l)}$$

$$\boldsymbol{q}_t,\ \boldsymbol{k}_t,\ \boldsymbol{v}_t = \boldsymbol{x}_t \boldsymbol{W}_Q,\ \boldsymbol{x}_t \boldsymbol{W}_K,\ \boldsymbol{x}_t \boldsymbol{W}_V$$

Key K
Value V
Query Q



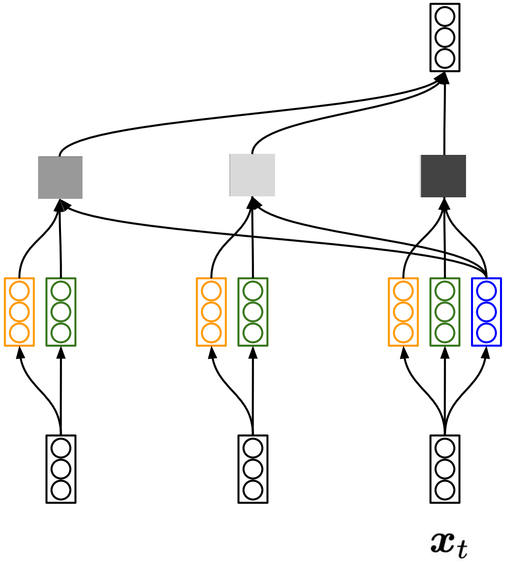
How do Transformers work?

This vector is a function of *all* previous inputs!

$$\boldsymbol{o}_t = \sum_{j=1}^{t} \frac{\exp(\boldsymbol{q}_t^\top \boldsymbol{k}_j)}{\sum_{l=1}^{t} \exp(\boldsymbol{q}_t^\top \boldsymbol{k}_l)} \boldsymbol{v}_j$$

$$\boldsymbol{q}_t,\ \boldsymbol{k}_t,\ \boldsymbol{v}_t = \boldsymbol{x}_t \boldsymbol{W}_Q,\ \boldsymbol{x}_t \boldsymbol{W}_K,\ \boldsymbol{x}_t \boldsymbol{W}_V$$

Key K
Value V
Query Q

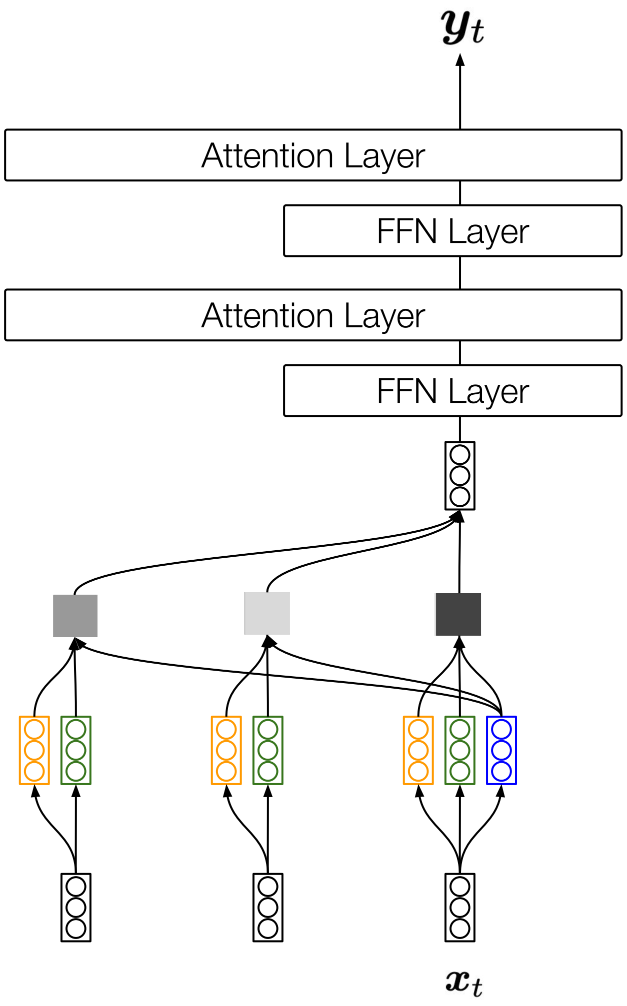


How do Transformers work?

$$\boldsymbol{o}_t = \sum_{j=1}^{t} \frac{\exp(\boldsymbol{q}_t^\top \boldsymbol{k}_j)}{\sum_{l=1}^{t} \exp(\boldsymbol{q}_t^\top \boldsymbol{k}_l)} \boldsymbol{v}_j$$

$$\boldsymbol{q}_t,\ \boldsymbol{k}_t,\ \boldsymbol{v}_t = \boldsymbol{x}_t \boldsymbol{W}_Q,\ \boldsymbol{x}_t \boldsymbol{W}_K,\ \boldsymbol{x}_t \boldsymbol{W}_V$$

Key K
Value V
Query Q

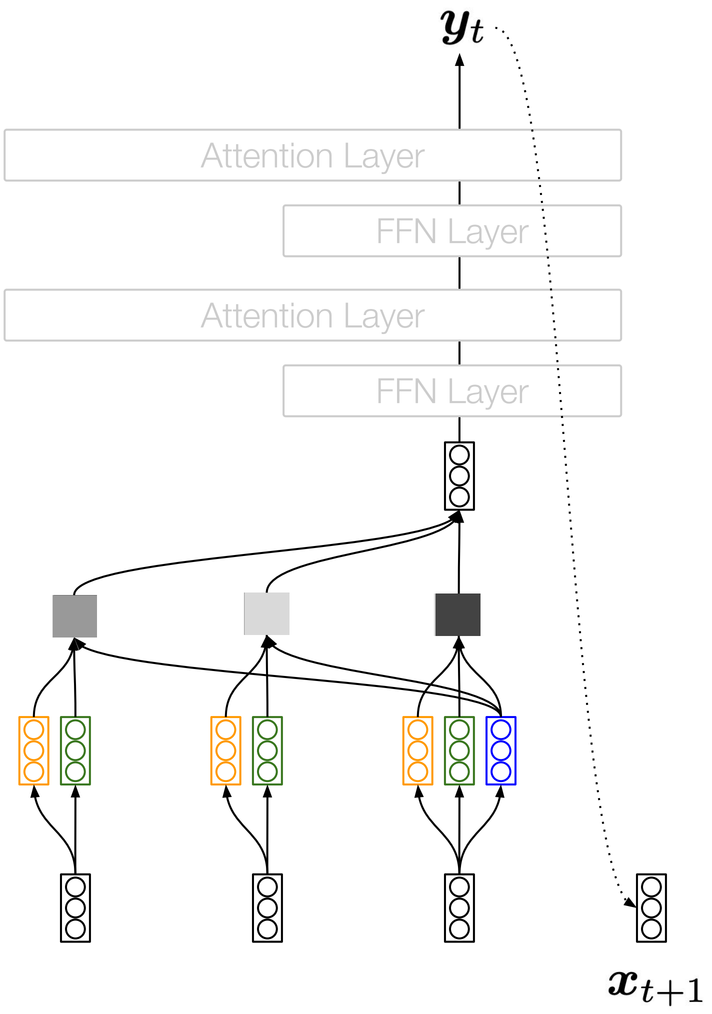


How do Transformers work?

$$\boldsymbol{o}_t = \sum_{j=1}^{t} \frac{\exp(\boldsymbol{q}_t^\top \boldsymbol{k}_j)}{\sum_{l=1}^{t} \exp(\boldsymbol{q}_t^\top \boldsymbol{k}_l)} \boldsymbol{v}_j$$

$$\boldsymbol{q}_t,\ \boldsymbol{k}_t,\ \boldsymbol{v}_t = \boldsymbol{x}_t \boldsymbol{W}_Q,\ \boldsymbol{x}_t \boldsymbol{W}_K,\ \boldsymbol{x}_t \boldsymbol{W}_V$$

Key K
Value V
Query Q

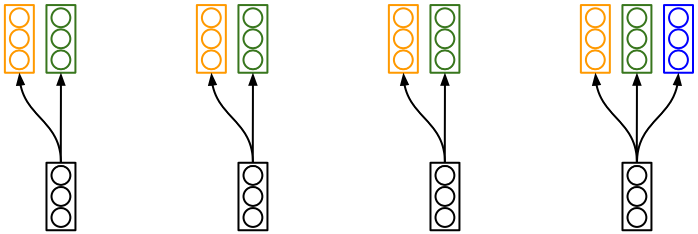


How do Transformers work?

$$\boldsymbol{o}_t = \sum_{j=1}^{t} \frac{\exp(\boldsymbol{q}_t^\top \boldsymbol{k}_j)}{\sum_{l=1}^{t} \exp(\boldsymbol{q}_t^\top \boldsymbol{k}_l)} \boldsymbol{v}_j$$

$$\boldsymbol{q}_t,\ \boldsymbol{k}_t,\ \boldsymbol{v}_t = \boldsymbol{x}_t \boldsymbol{W}_Q,\ \boldsymbol{x}_t \boldsymbol{W}_K,\ \boldsymbol{x}_t \boldsymbol{W}_V$$

Key K
Value V
Query Q

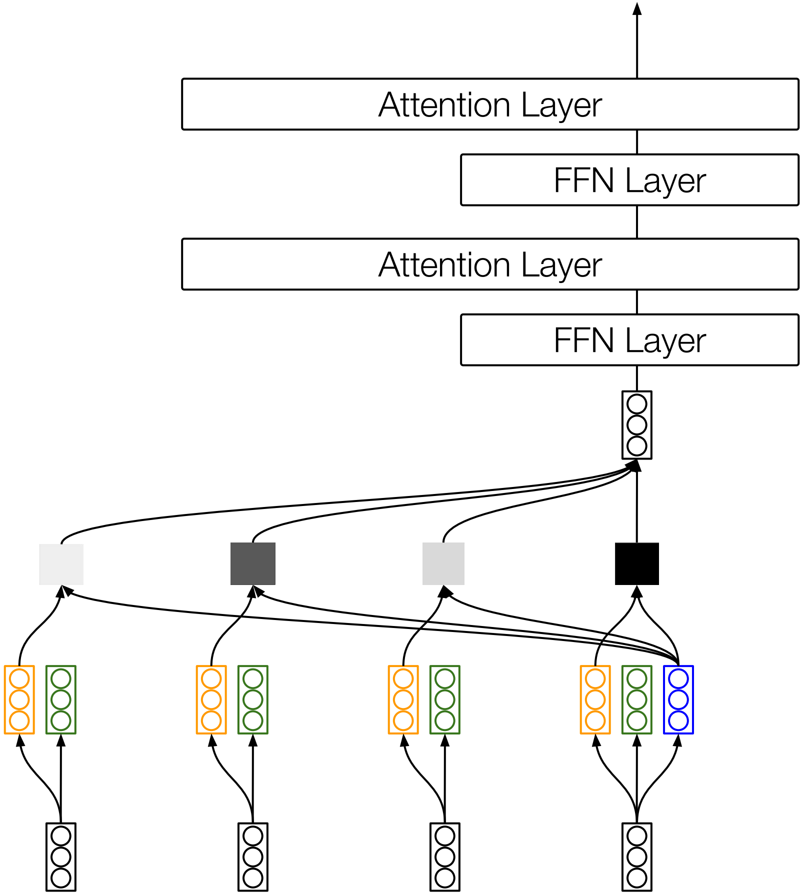


How do Transformers work?

$$o_t = \sum_{j=1}^{t} \frac{\exp(q_t^\top k_j)}{\sum_{l=1}^{t} \exp(q_t^\top k_l)} v_j$$

$$q_t,\ k_t,\ v_t = x_t W_Q,\ x_t W_K,\ x_t W_V$$

Key K
Value V
Query Q

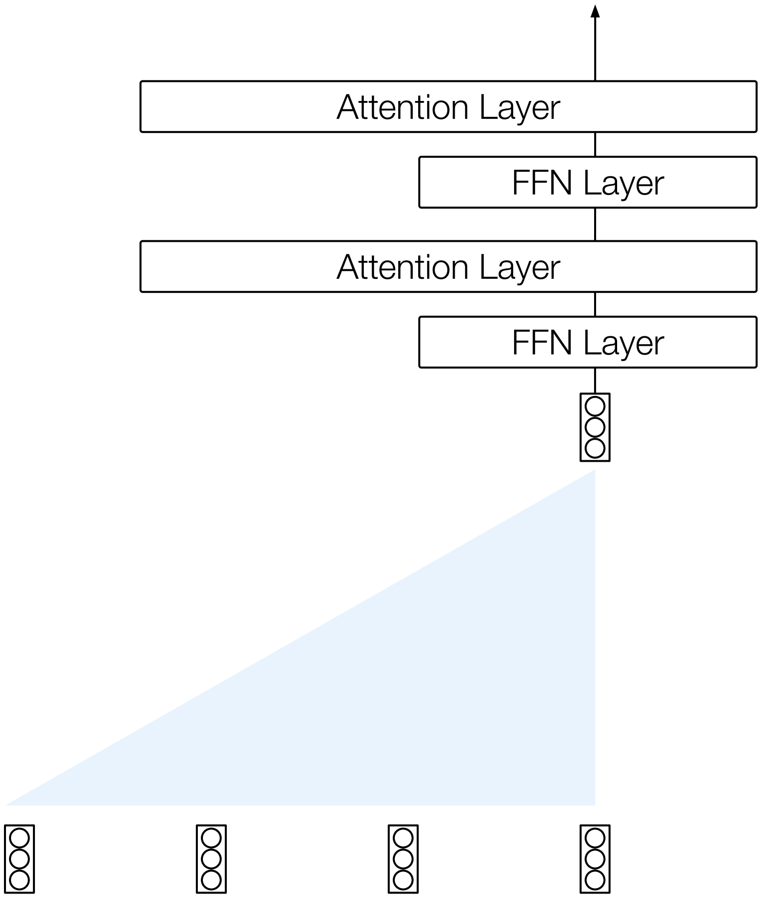


How do Transformers work?

$$\boldsymbol{o}_t = \sum_{j=1}^{t} \frac{\exp(\boldsymbol{q}_t^\top \boldsymbol{k}_j)}{\sum_{l=1}^{t} \exp(\boldsymbol{q}_t^\top \boldsymbol{k}_l)} \boldsymbol{v}_j$$

$$\boldsymbol{q}_t,\ \boldsymbol{k}_t,\ \boldsymbol{v}_t = \boldsymbol{x}_t \boldsymbol{W}_Q,\ \boldsymbol{x}_t \boldsymbol{W}_K,\ \boldsymbol{x}_t \boldsymbol{W}_V$$

Key K
Value V
Query Q

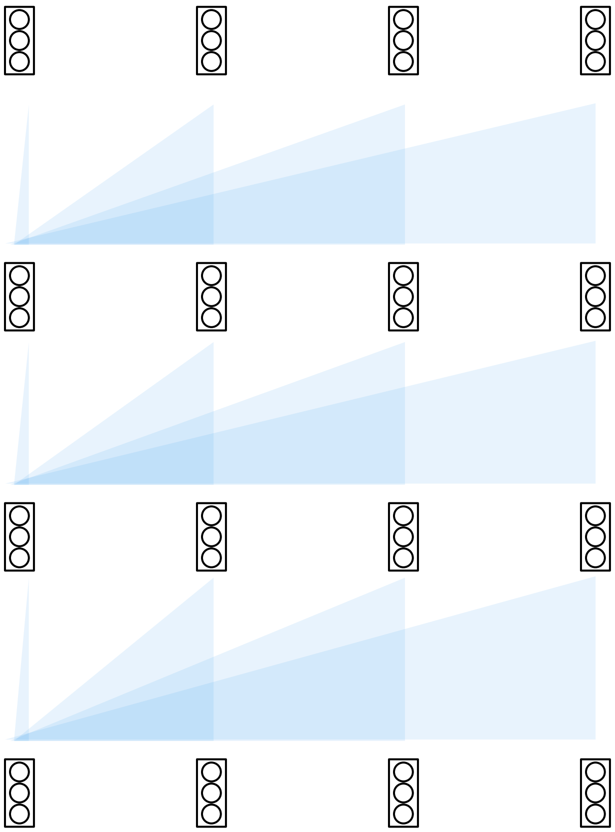


How do Transformers work?

Transformers use flexible but uninterpretable internal operations to compute the output given the input.

$$\boldsymbol{q}_t,\ \boldsymbol{k}_t,\ \boldsymbol{v}_t = \boldsymbol{x}_t\boldsymbol{W}_Q,\ \boldsymbol{x}_t\boldsymbol{W}_K,\ \boldsymbol{x}_t\boldsymbol{W}_V$$

Can we implement “hard” rules in Transformers?



Implementing Logic in Transformers

A Logic for Expressing Log-Precision Transformers

William Merrill
New York University
willm@nyu.edu

Ashish Sabharwal
Allen Institute for AI
ashishs@allenai.org

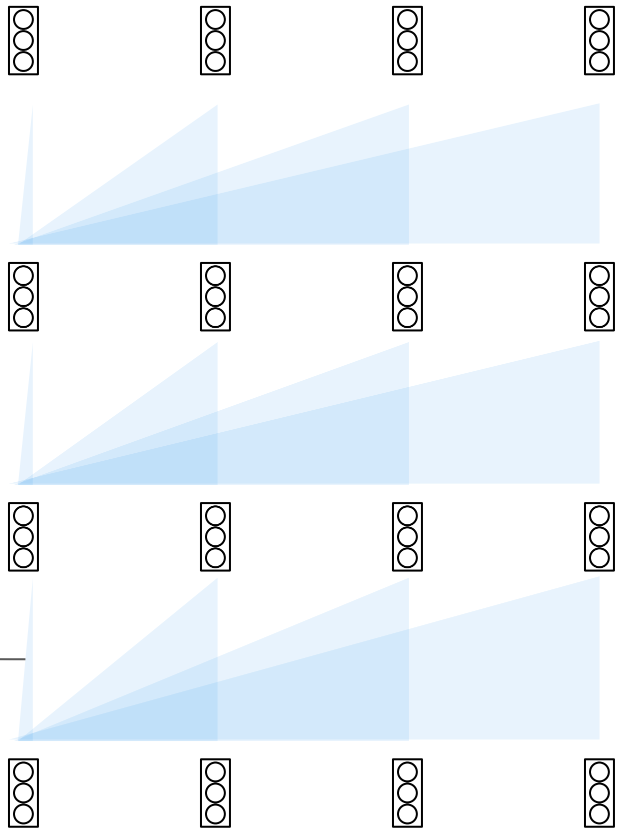
Any log-precision transformer can be re-expressed as a sentence in $\mathsf{FO}(\mathsf{M})$ logic, e.g.:

$$\mathsf{M}i.\ \mathtt{a}(i)\ \wedge\ \mathsf{M}j.\ \mathtt{b}(j)\ \wedge\ \neg\exists k,\ell.\ (\mathtt{a}(k) \wedge \mathtt{b}(\ell) \wedge \ell < k)$$

(m a's followed by m b's, i.e., $a^m b^m$)

aaaabbbb ✓ aaabbbbb ✗ baaaabbb ✗

Implementing Logic in Transformers



Any log-precision transformer can be re-expressed as a sentence in FO(M) logic, e.g.:

$$\mathsf{M}i.\ \mathtt{a}(i)\ \wedge\ \mathsf{M}j.\ \mathtt{b}(j)\ \wedge\ \neg\exists k,\ell.\ (\mathtt{a}(k) \wedge \mathtt{b}(\ell) \wedge \ell < k)$$

(m $\mathtt{a}$'s followed by m $\mathtt{b}$'s, i.e., $\mathtt{a}^m\mathtt{b}^m$)

aaaabbbb ✓ aaabbbbb ✗ baaaabbb ✗

Trained Transformer weights can be "read out" in the form of logical rules → more interpretable!

Implementing Logic in Transformers

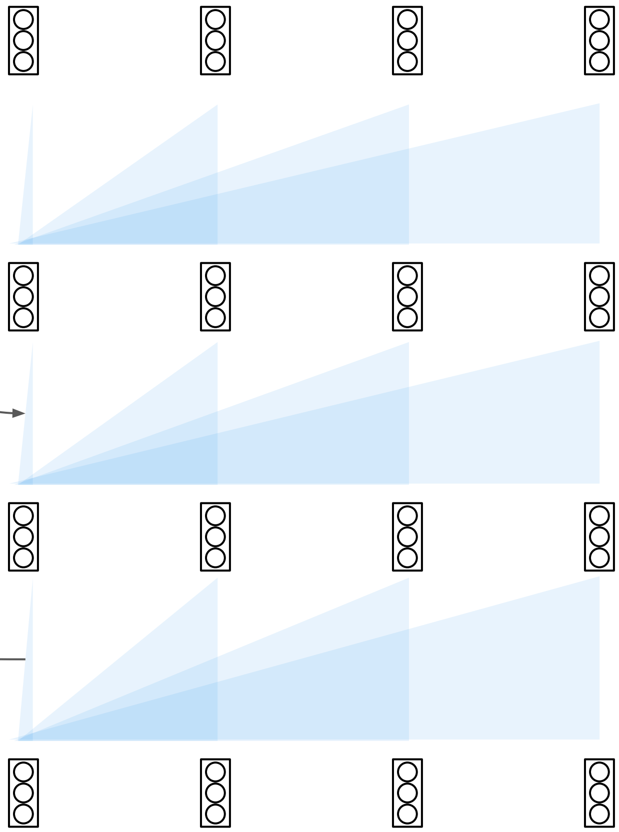
These logical rules can be "compiled" into the weights of Transformers → more controllable!

Any log-precision transformer can be re-expressed as a sentence in FO(M) logic, e.g.:

$$\mathsf{M}i.\ \mathtt{a}(i)\ \wedge\ \mathsf{M}j.\ \mathtt{b}(j)\ \wedge\ \neg\exists k,\ell.\ (\mathtt{a}(k) \wedge \mathtt{b}(\ell) \wedge \ell < k)$$

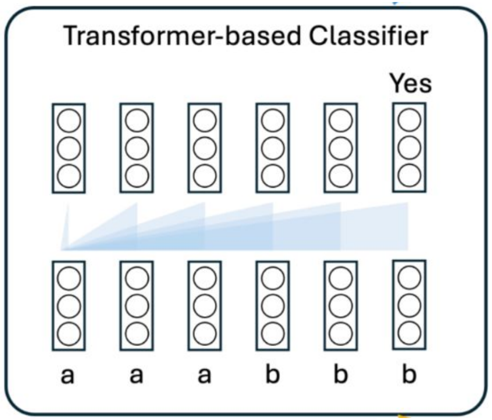
(m a's followed by m b's, i.e., $a^m b^m$)

aaaabbbb ✓ aaabbbbb ✗ baaaabbb ✗

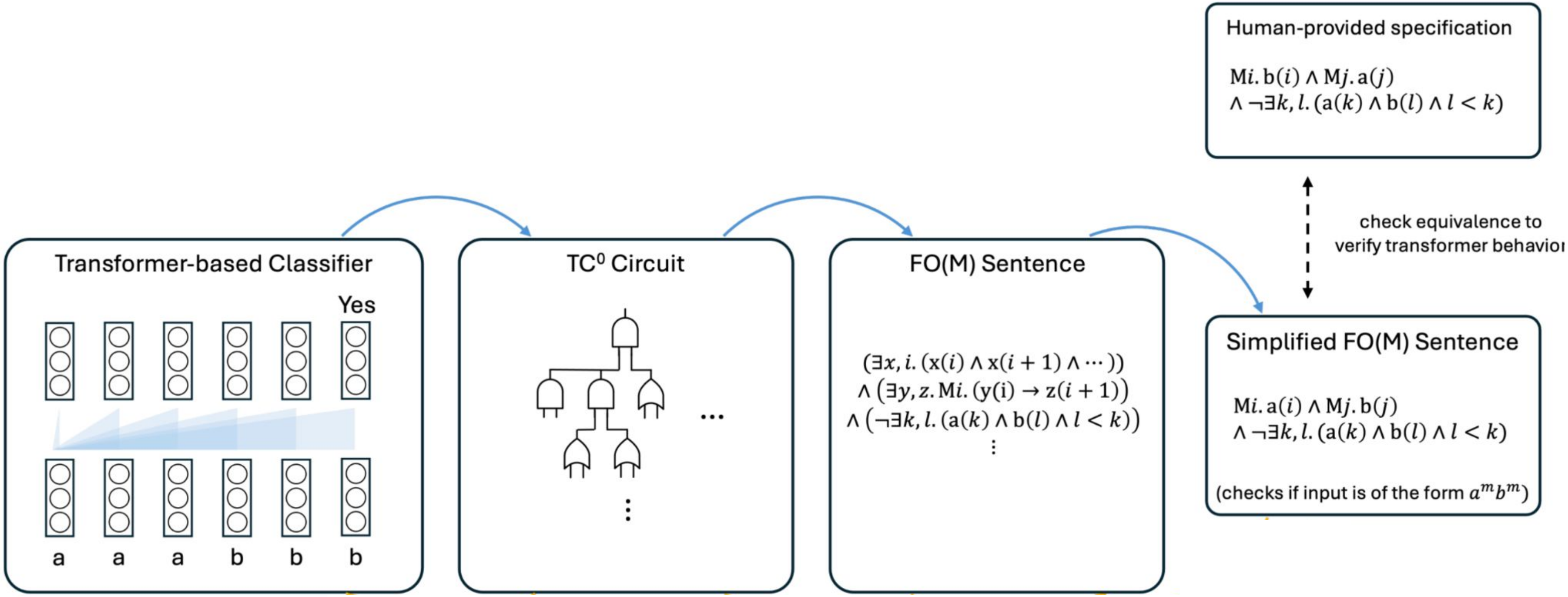


Trained Transformer weights can be "read out" in the form of logical rules → more interpretable!

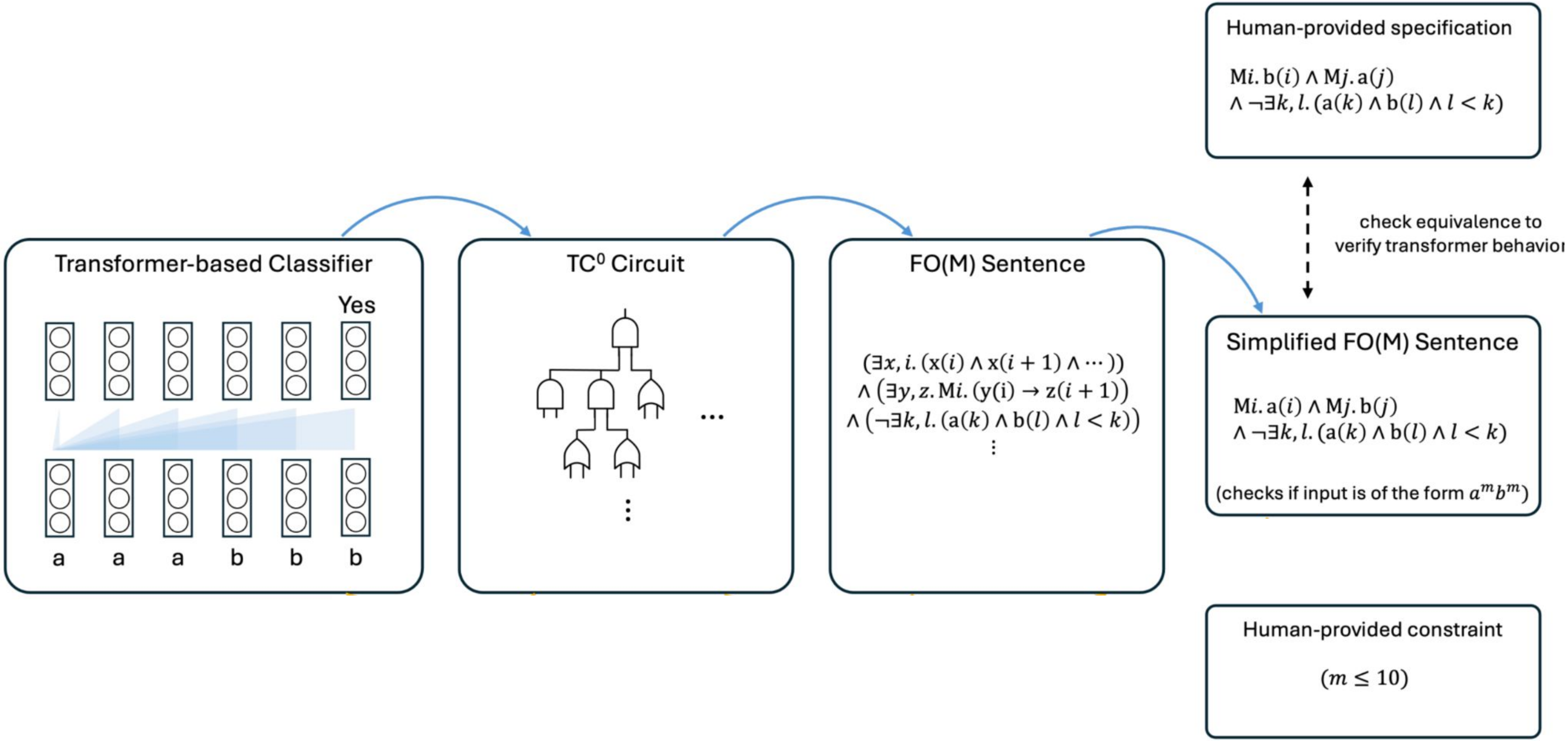
Our project



Our project

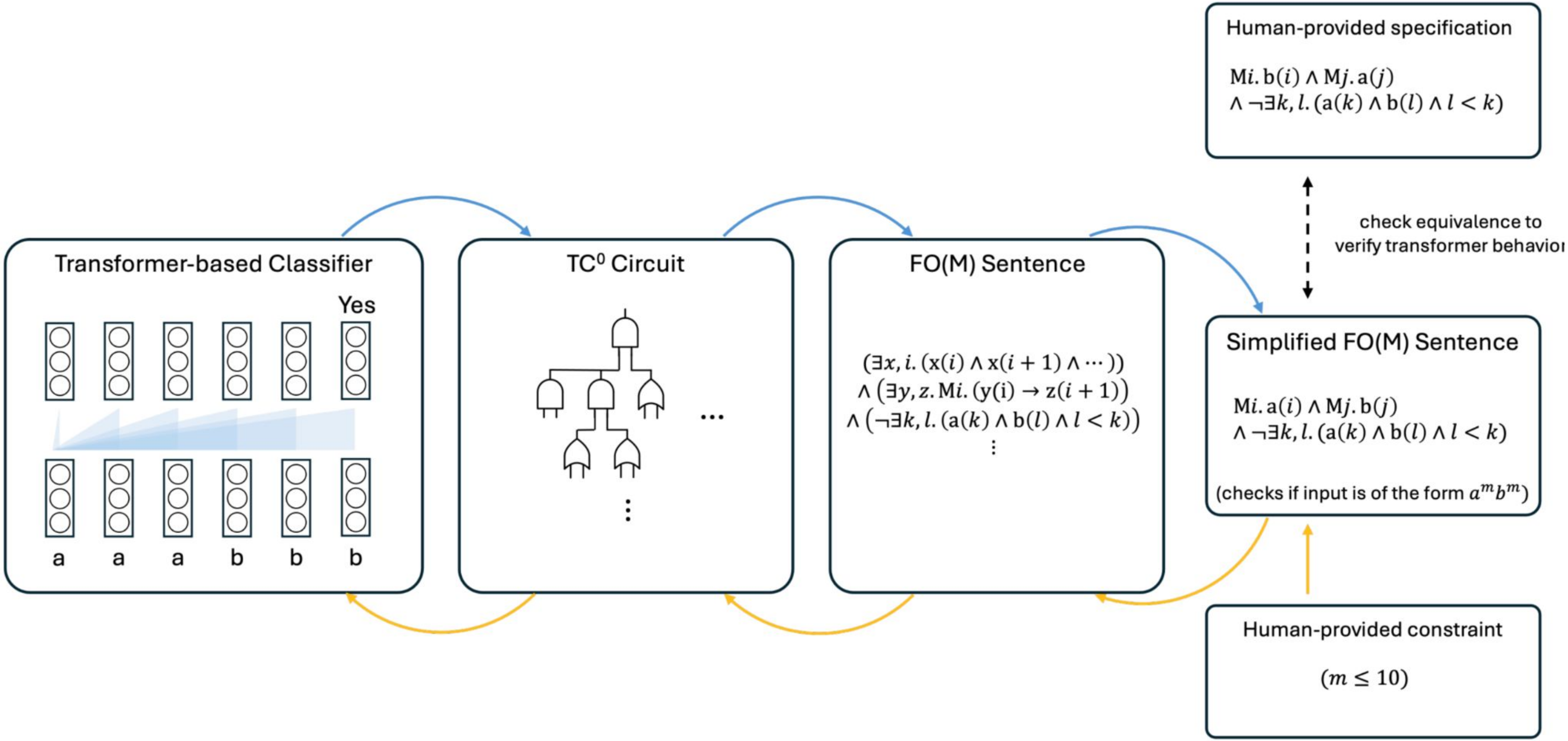

Human-provided specification
$\mathrm{M}i.\,\mathrm{b}(i) \wedge \mathrm{M}j.\,\mathrm{a}(j)$
$\wedge \neg\exists k, l.\,(\mathrm{a}(k) \wedge \mathrm{b}(l) \wedge l < k)$
check equivalence to
verify transformer behavior
Transformer-based Classifier
Yes
a
a
a
b
b
b
TC^0 Circuit
...
FO(M) Sentence
$(\exists x, i.\,(\mathrm{x}(i) \wedge \mathrm{x}(i+1) \wedge \cdots))$
$\wedge\,(\exists y, z.\,\mathrm{M}i.\,(\mathrm{y}(\mathrm{i}) \rightarrow \mathrm{z}(i+1)))$
$\wedge\,(\neg\exists k, l.\,(\mathrm{a}(k) \wedge \mathrm{b}(l) \wedge l < k))$
Simplified FO(M) Sentence
$\mathrm{M}i.\,\mathrm{a}(i) \wedge \mathrm{M}j.\,\mathrm{b}(j)$
$\wedge \neg\exists k, l.\,(\mathrm{a}(k) \wedge \mathrm{b}(l) \wedge l < k)$
(checks if input is of the form $a^m b^m$)

Our project



Transformer-based Classifier
Yes
a a a b b b
TC^0 Circuit
...
FO(M) Sentence
$(\exists x, i. (x(i) \wedge x(i+1) \wedge \cdots))$
$\wedge (\exists y, z. \mathrm{M}i. (y(i) \rightarrow z(i+1)))$
$\wedge (\neg\exists k, l. (a(k) \wedge b(l) \wedge l < k))$
Human-provided specification
$\mathrm{M}i. b(i) \wedge \mathrm{M}j. a(j)$
$\wedge \neg\exists k, l. (a(k) \wedge b(l) \wedge l < k)$
check equivalence to
verify transformer behavior
Simplified FO(M) Sentence
$\mathrm{M}i. a(i) \wedge \mathrm{M}j. b(j)$
$\wedge \neg\exists k, l. (a(k) \wedge b(l) \wedge l < k)$
(checks if input is of the form $a^m b^m$)
Human-provided constraint
$(m \leq 10)$

Our project


Transformer-based Classifier
Yes
a a a b b b
TC^0 Circuit
...
FO(M) Sentence
$(\exists x, i.\,(x(i) \land x(i+1) \land \cdots))$
$\land\,(\exists y, z.\,\mathrm{M}i.\,(y(i) \to z(i+1)))$
$\land\,(\neg\exists k, l.\,(a(k) \land b(l) \land l < k))$
Human-provided specification
$\mathrm{M}i.\,b(i) \land \mathrm{M}j.\,a(j)$
$\land\,\neg\exists k, l.\,(a(k) \land b(l) \land l < k)$
check equivalence to
verify transformer behavior
Simplified FO(M) Sentence
$\mathrm{M}i.\,a(i) \land \mathrm{M}j.\,b(j)$
$\land\,\neg\exists k, l.\,(a(k) \land b(l) \land l < k)$
(checks if input is of the form $a^m b^m$)
Human-provided constraint
$(m \leq 10)$

Ongoing Work