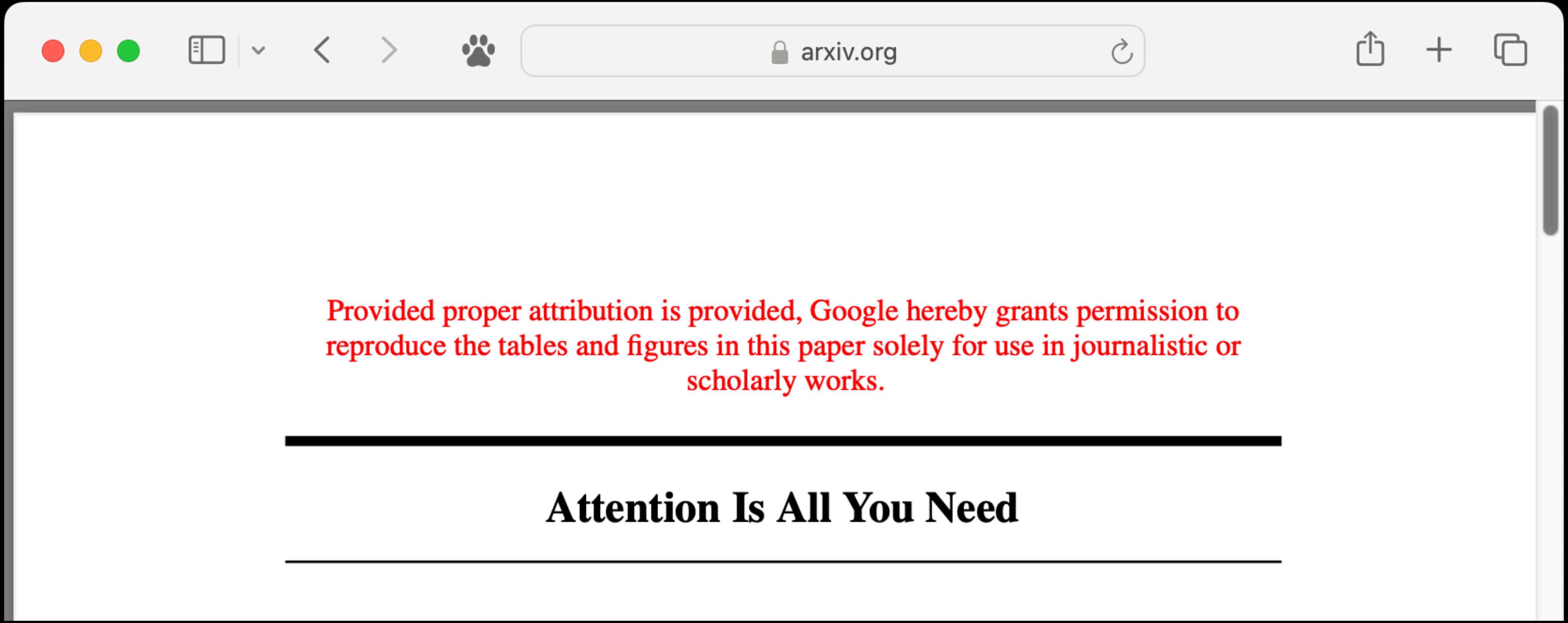
Transformers

a social history

NLU Lab, 3 April 2024

1. What a Transformer is
2. Why they are incredibly popular
3. Where they are used
4. Problems & Solutions

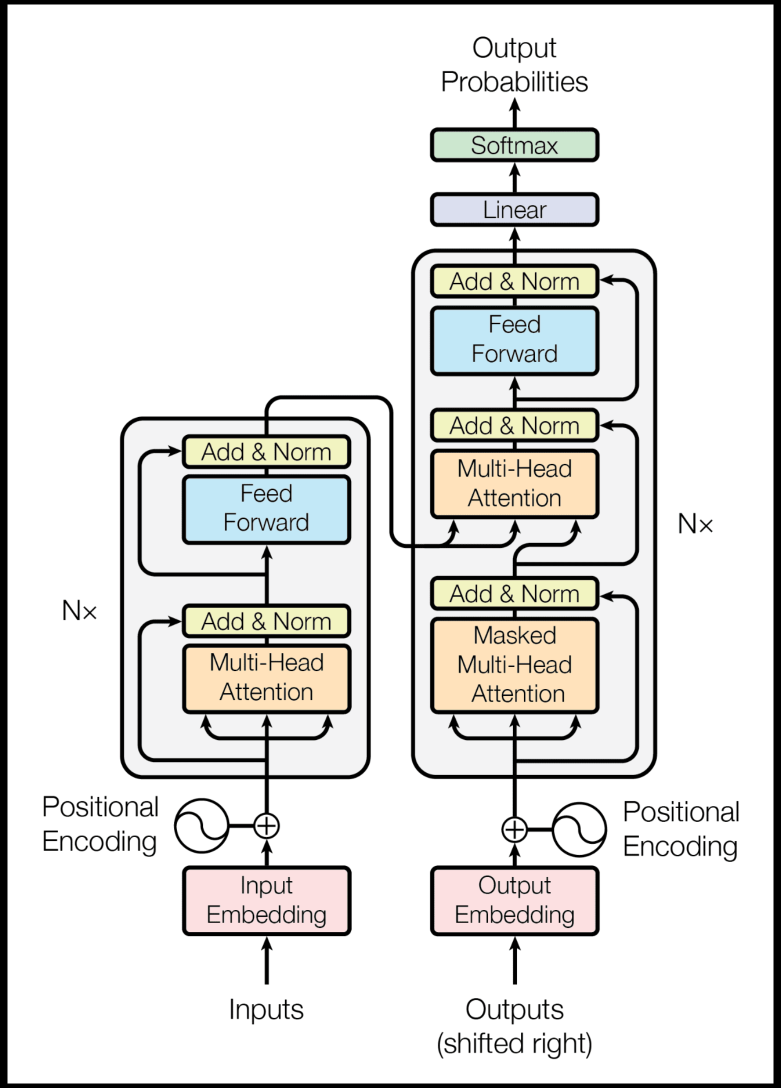
Whence the Transformer



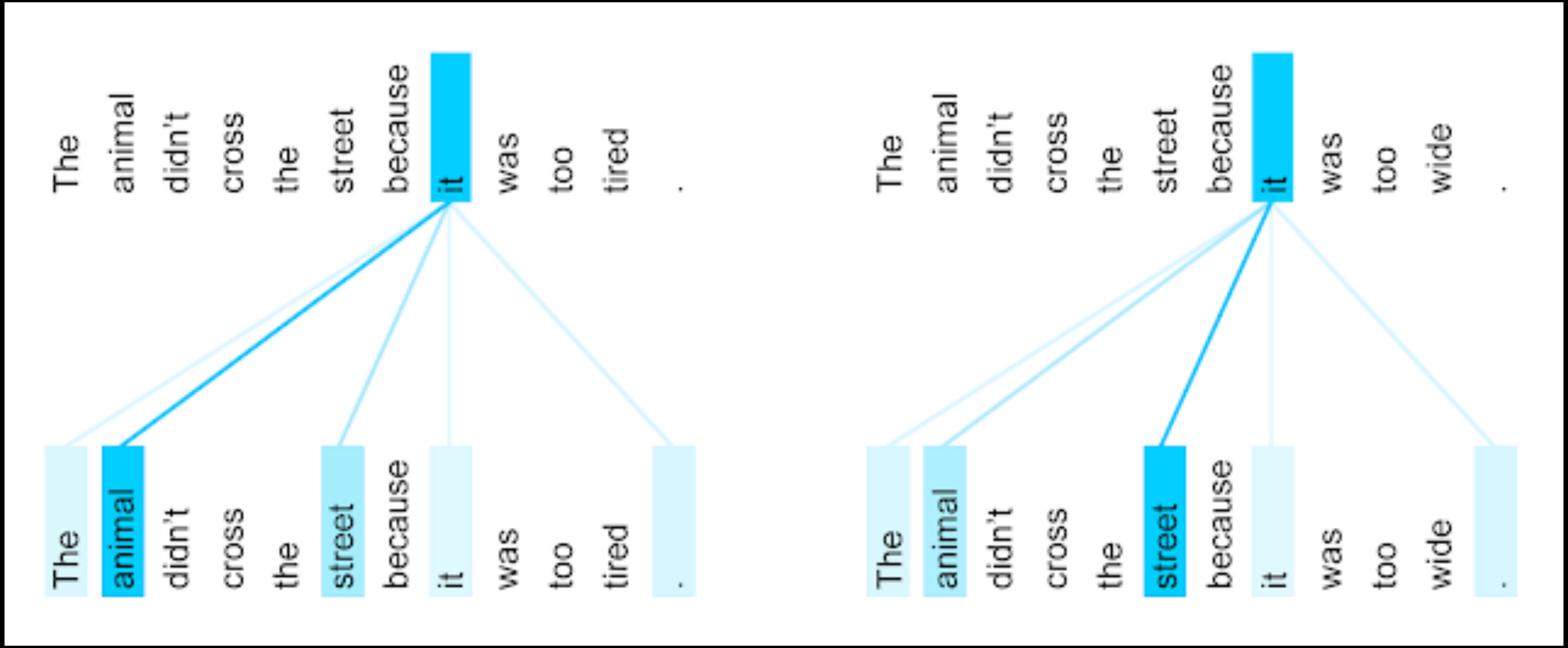
Provided proper attribution is provided, Google hereby grants permission to reproduce the tables and figures in this paper solely for use in journalistic or scholarly works.

Attention Is All You Need

Transformer Architecture



Attention, visualized



The animal didn't cross the street because it was too tired .
The animal didn't cross the street because it was too tired .
The animal didn't cross the street because it was too wide .
The animal didn't cross the street because it was too wide .

Why were transformers so revolutionary?

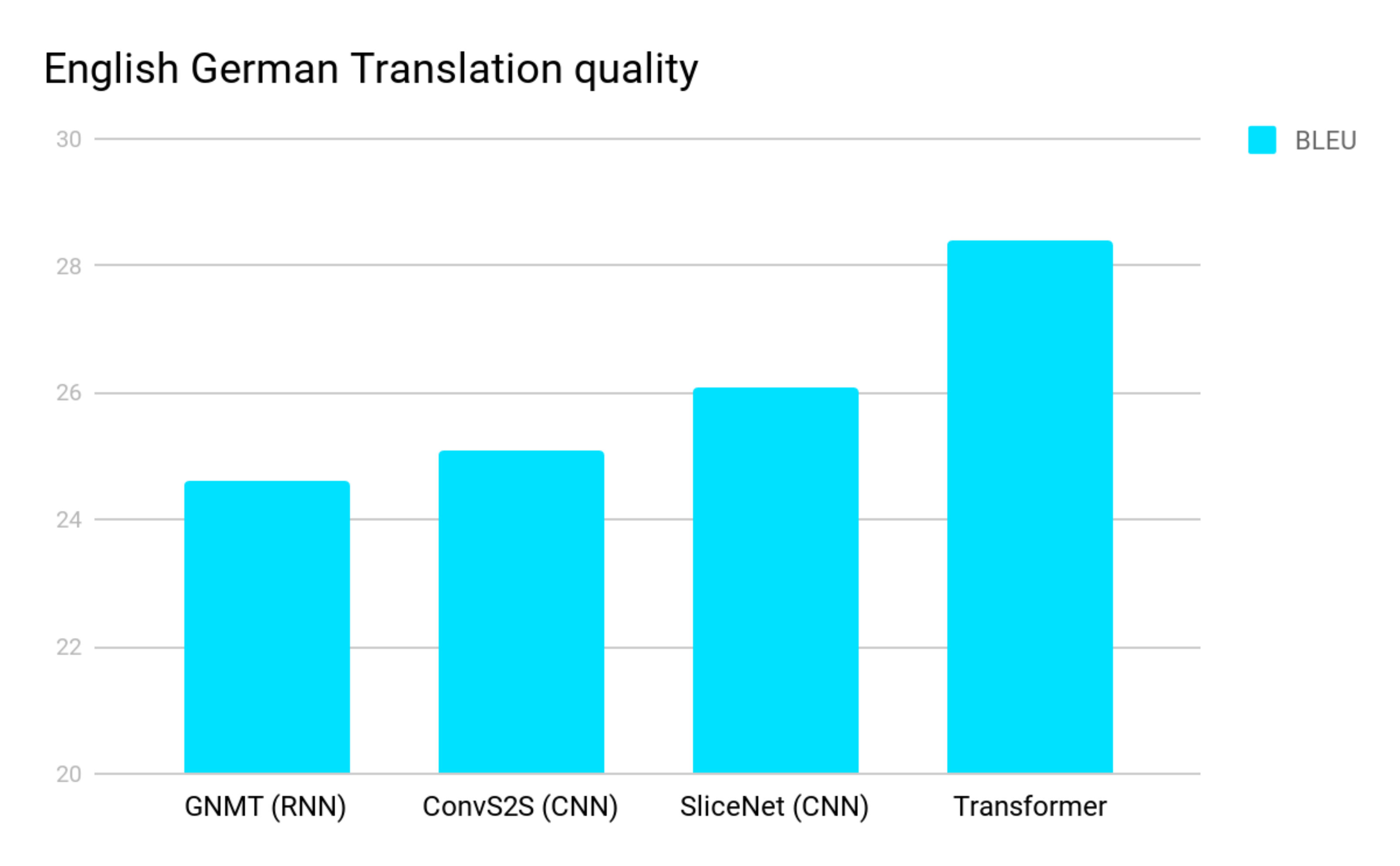
Parallel Training

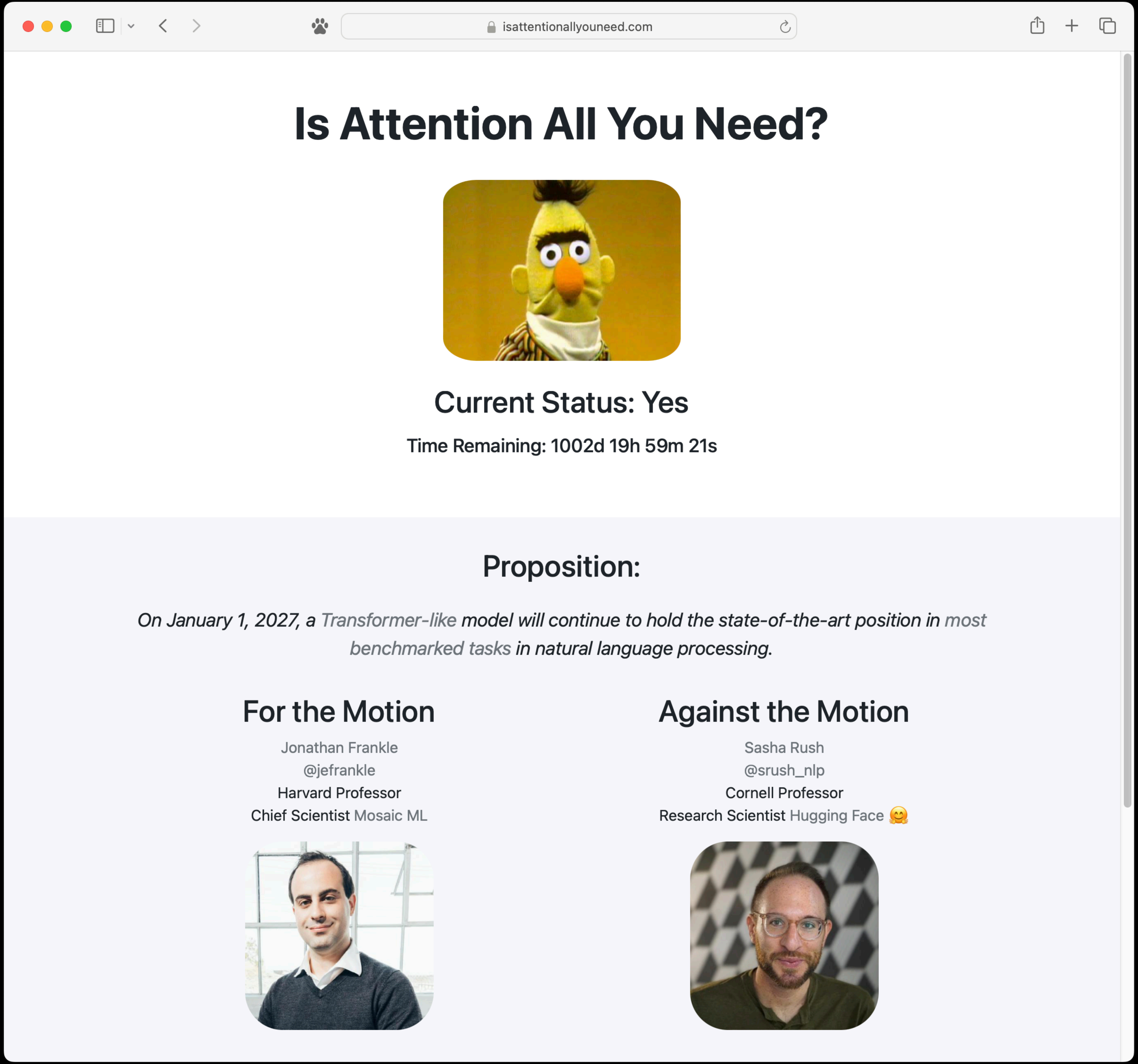
Recurrent Models have an inherent ... recurrence in their sequence dimension

→ training time is proportional to sequence length

Transformers: No recurrence = parallelism!

Performance!





Is Attention All You Need?

Current Status: Yes

Time Remaining: 1002d 19h 59m 21s

Proposition:

On January 1, 2027, a Transformer-like model will continue to hold the state-of-the-art position in most benchmarked tasks in natural language processing.

For the Motion

Jonathan Frankle
@jefrankle
Harvard Professor
Chief Scientist Mosaic ML

Against the Motion

Sasha Rush
@srush_nlp
Cornell Professor
Research Scientist Hugging Face 🤗

Transformers: Where are they now?

Basically everywhere

GPT (ChatGPT) [Decoder-only]

Claude, Gemini (probably?)

BERT, Pythia, OLMo, T5 (Open source!)

Problems with Transformers

Inference Cost

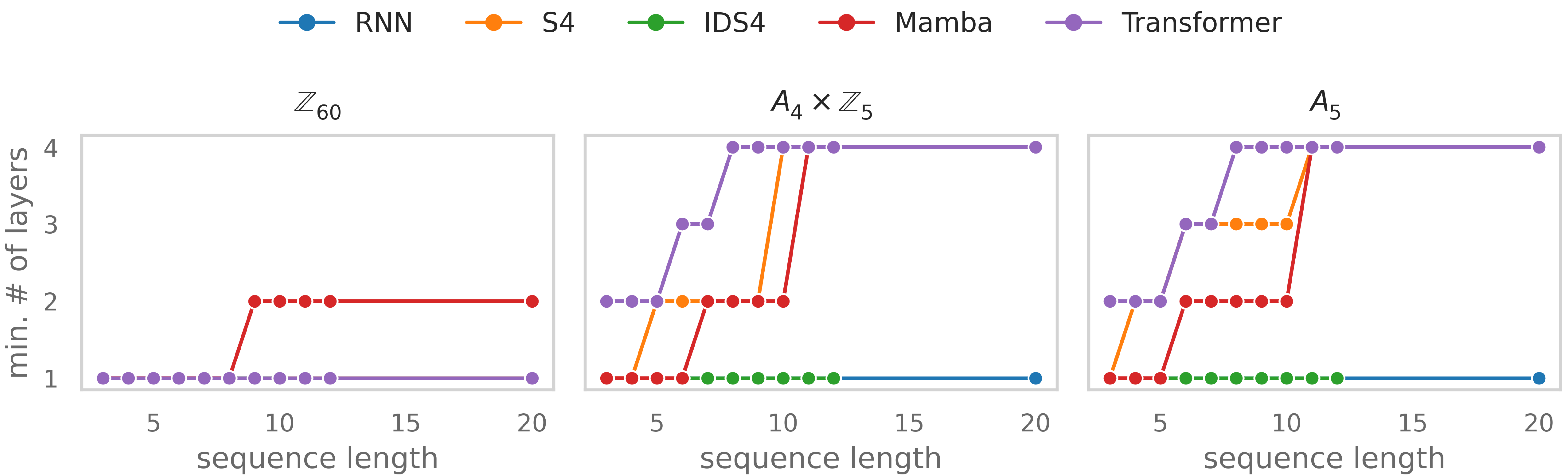
- Transformers are parallel @ training time, but autoregressive at inference
- Attention is expensive: Quadratic complexity in sequence length
- RNNs are actually better here!
- $$$ for long generations

Length Generalization

- Transformers don't have any inherent notion of sequential position
- Traditional positional encodings don't seem to yield good length generalizations
- Context lengths are getting quite long these days...

Representational Power

- Transformers are actually *weaker* than RNNs, in terms of what kind of functions they can learn to express

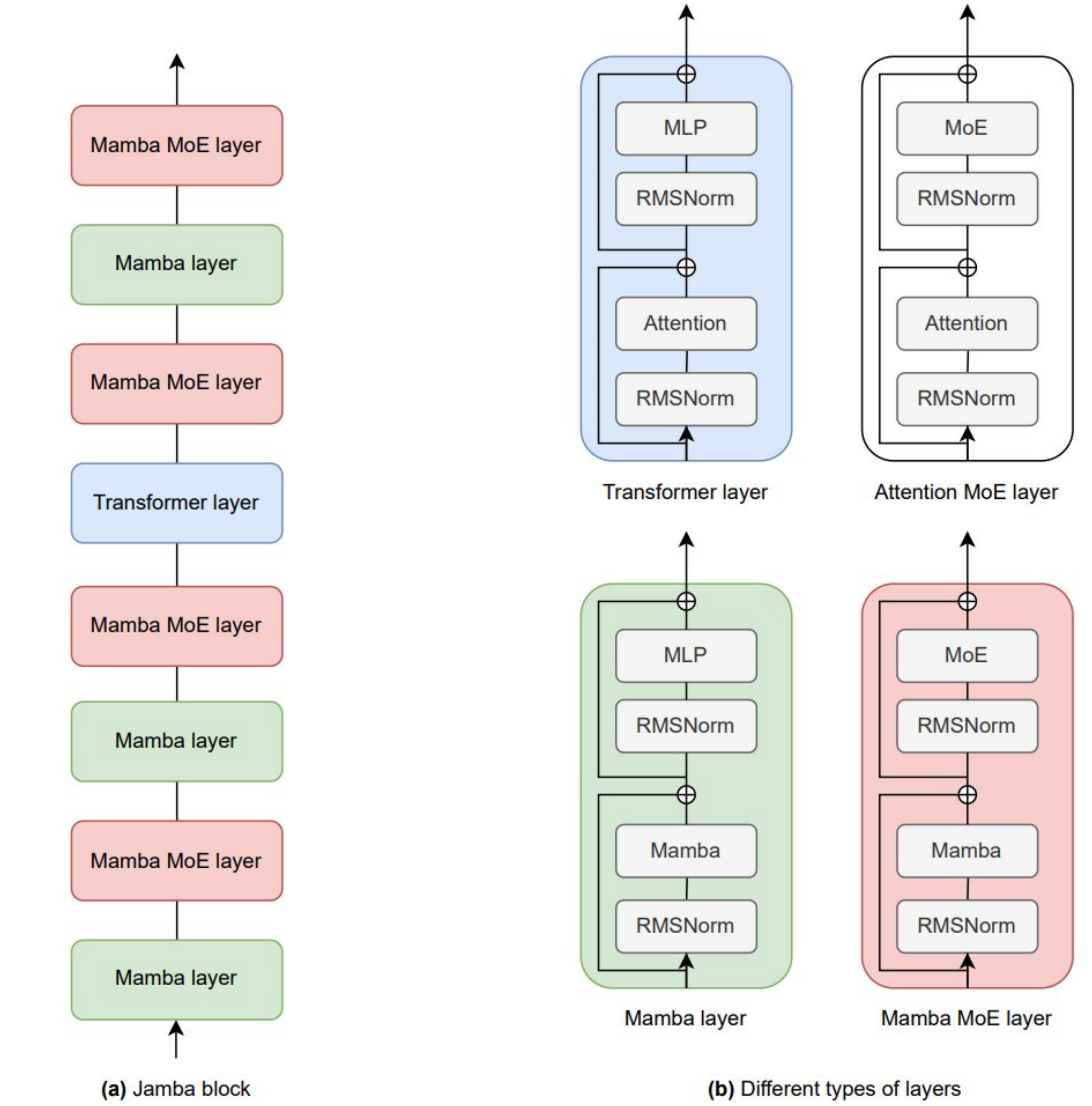

RNN
S4
IDS4
Mamba
Transformer
$\mathbb{Z}_{60}$
$A_4 \times \mathbb{Z}_5$
A_5
min. # of layers
1
2
3
4
5
10
15
20
sequence length

Solutions?

- Simplify attention mechanism to sub-quadratic complexity?
  - Sliding windows, approximations, etc
- New kinds of positional encodings
  - ROPE, Alibi
- New model architectures?
  - "Linear State Space models" ("parallelized RNNs" like Mamba), but see Merrill, Petty, Sabharwal (forthcoming)

Kitchen Sink Model? (Jamba, ~6 days old)

7:1 ratio of Mamba : Transformer layers



(a) Jamba block

(b) Different types of layers

Figure 1: **(a)** A single Jamba block. **(b)** Different types of layers. The implementation shown here is with $l = 8$, $a : m = 1 : 7$ ratio of attention-to-Mamba layers, and MoE applied every $e = 2$ layers.

DIRECTED BY

MICHAEL BAY