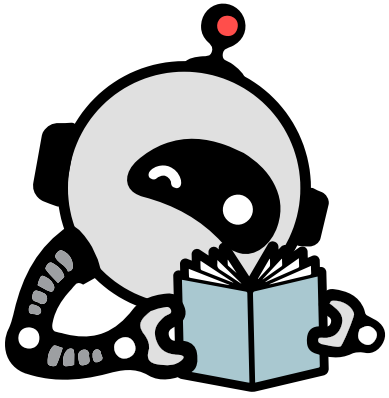
PROGRESS MEASURES FOR GROKKING VIA MECHANISTIC INTERPRETABILITY

L NANDA, LAWRENCE CHAN, TOM LIEBERUM, JESS SMITH, JA
STEINHARDT

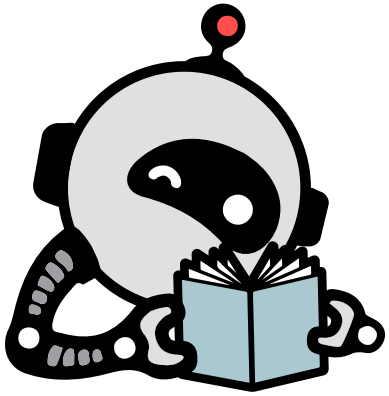
INTERNATIONAL CONFERENCE ON LEARING REPRESENTATIONS (ICLR) 2023

OUTLINE



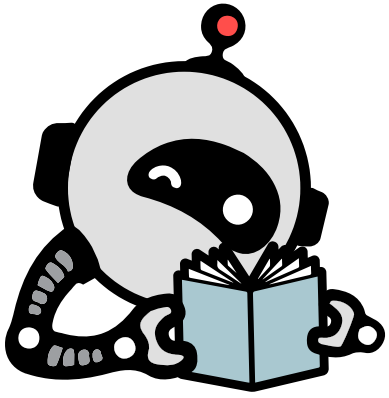
OUTLINE

- Mechanistic Interpretability



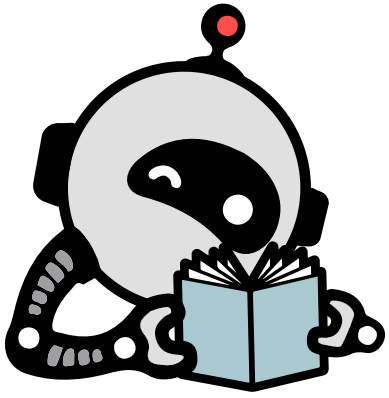
OUTLINE

- Mechanistic Interpretability
- grok /grŏk/



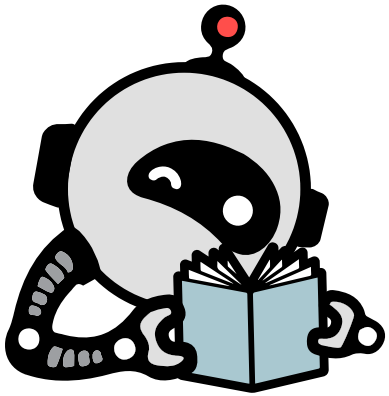
OUTLINE

- Mechanistic Interpretability
- grok /grŏk/
- Progress Measures



"Geometrically Guided Saliency Maps" Md Mahfuzur Rahman, Noah Lewis, Sergey Plis ICLR 2022 Workshop PAIR2Struct

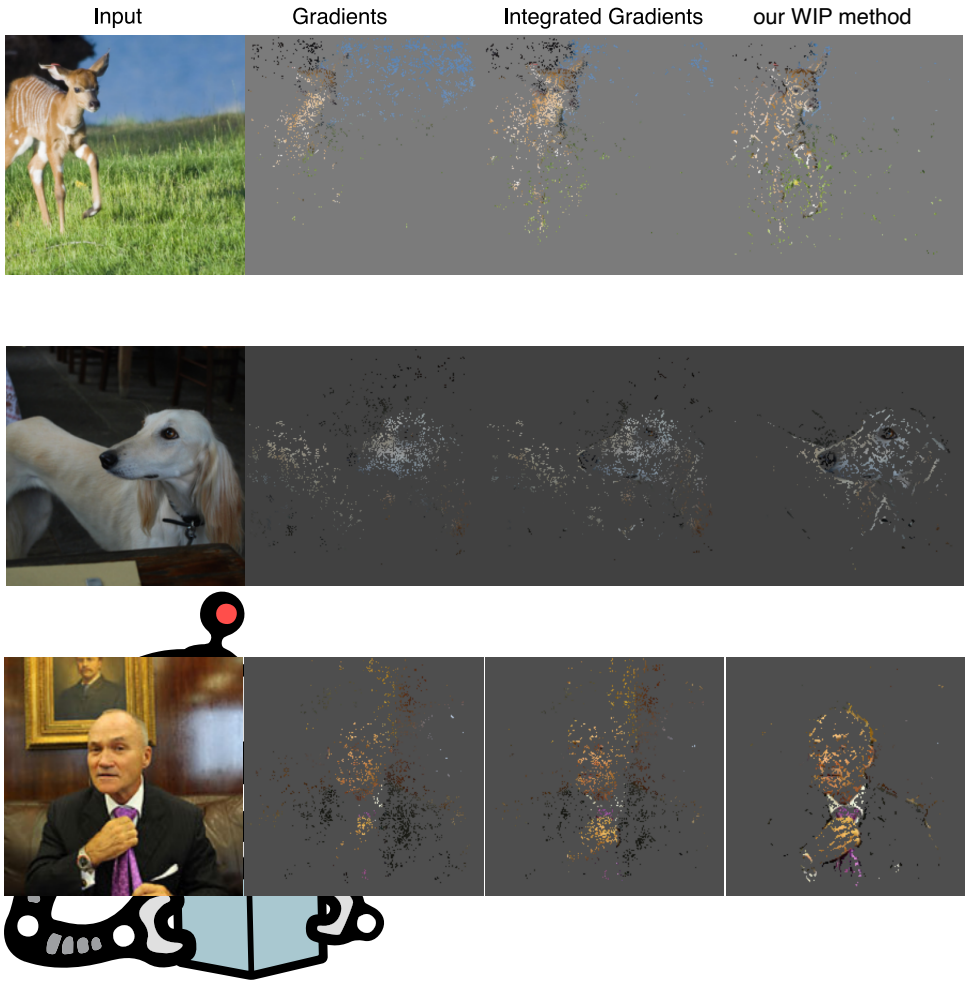
MECHANISTIC INTERPRETABILITY



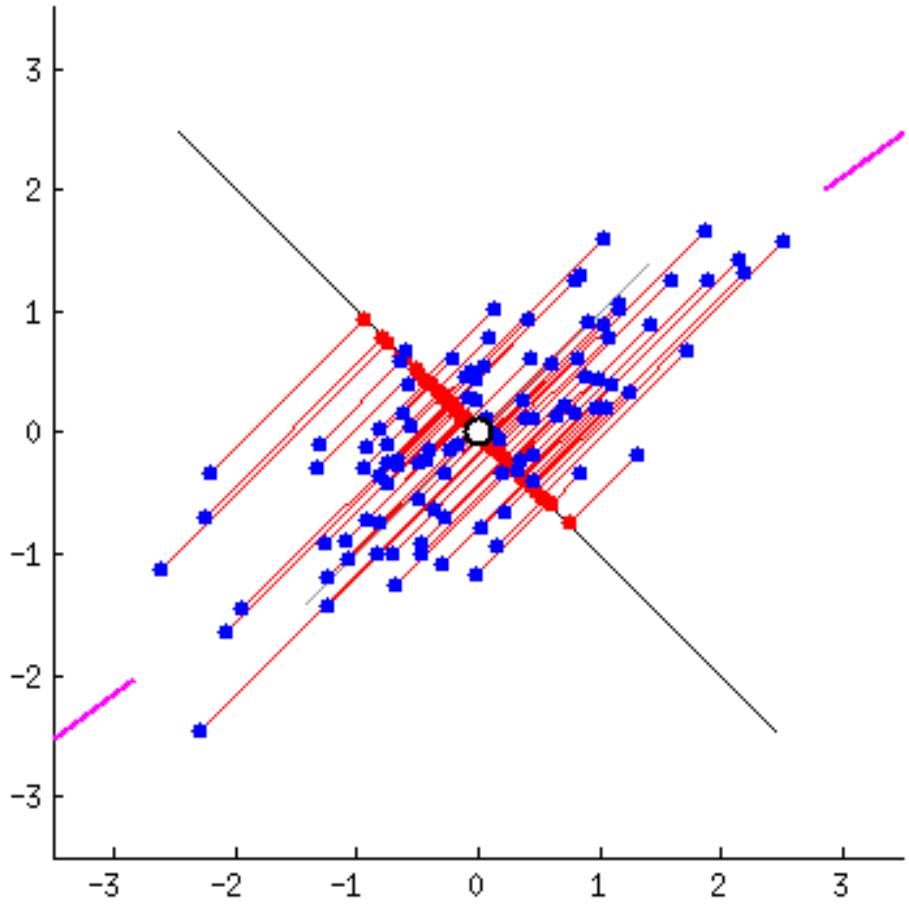
"Geometrically Guided Saliency Maps" Md Mahfuzur Rahman, Noah Lewis, Sergey Plis ICLR 2022 Workshop PAIR2Struct

TWO TYPES OF EXPLAINABLE AI

EXPLAIN THE DATA



INTERPRET THE MODEL



"Geometrically Guided Saliency Maps" Md Mahfuzur Rahman, Noah Lewis, Sergey Plis ICLR 2022 Workshop PAIR2Struct

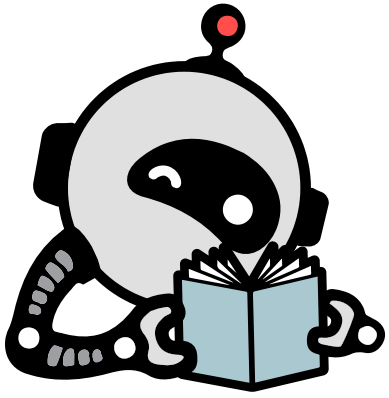
HOMEWORK PROBLEM

Implement a recurrent neural network which implements binary addition. The inputs are given as binary sequences, starting with the least significant binary digit.

$$100111 + 110010 = 1011001$$

would be represented as :

- Input 1: 1, 1, 1, 0, 0, 1, 0
- Input 2: 0, 1, 0, 0, 1, 1, 0
- Correct output: 1, 0, 0, 1, 1, 0, 1



HOMEWORK PROBLEM

Implement a recurrent neural network which implements binary addition. The inputs are given as binary sequences, starting with the least significant binary digit.

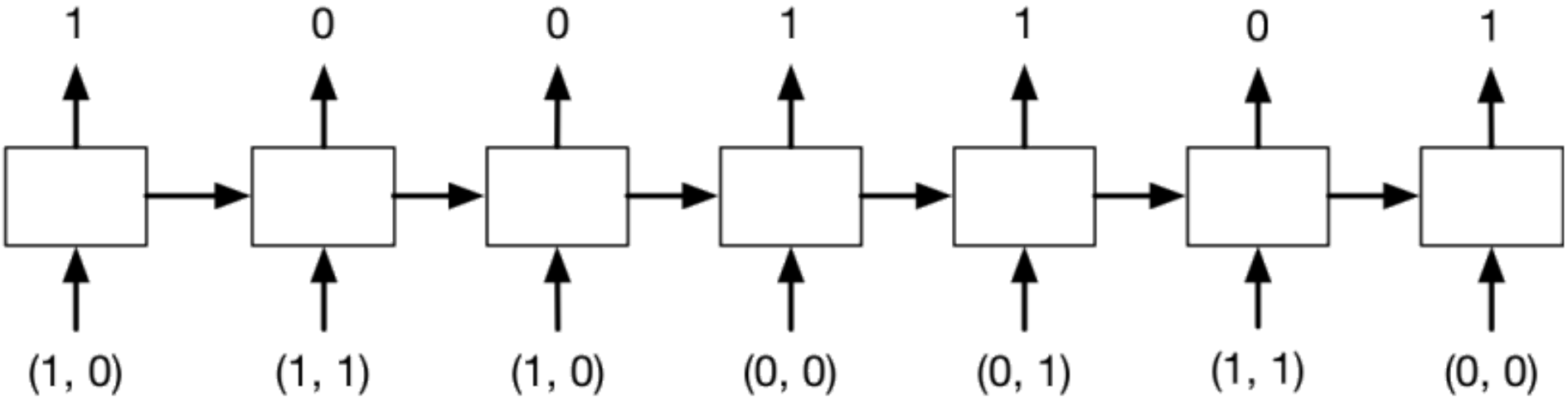
$$100111 + 110010 = 1011001$$

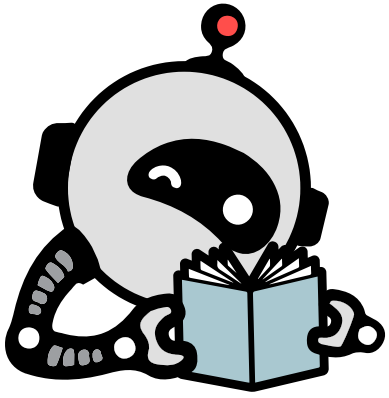
would be represented as :

- Input 1: 1, 1, 1, 0, 0, 1, 0
- Input 2: 0, 1, 0, 0, 1, 1, 0
- Correct output: 1, 0, 0, 1, 1, 0, 1

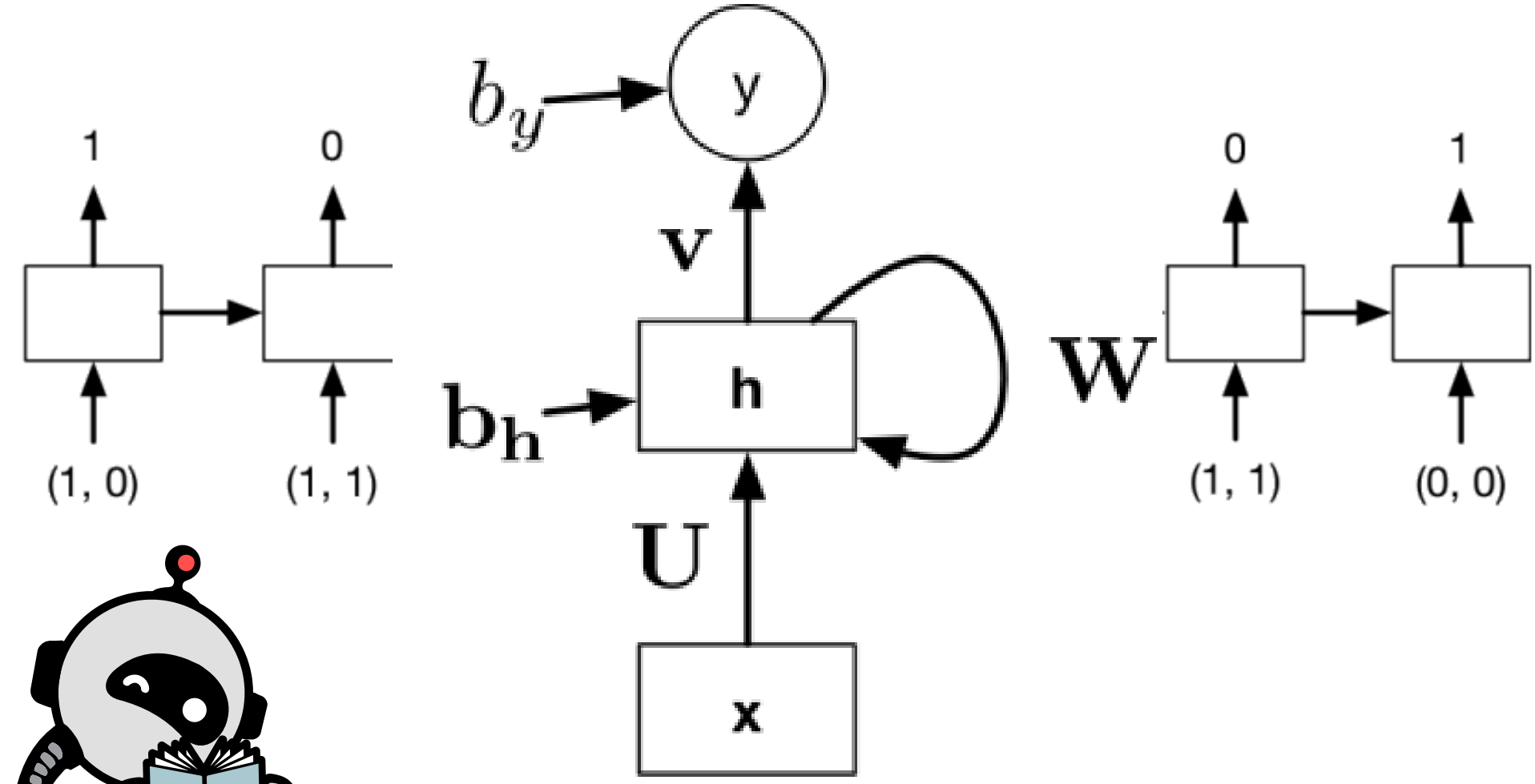
Set the parameter values manually - no training!

HOMEWORK PROBLEM



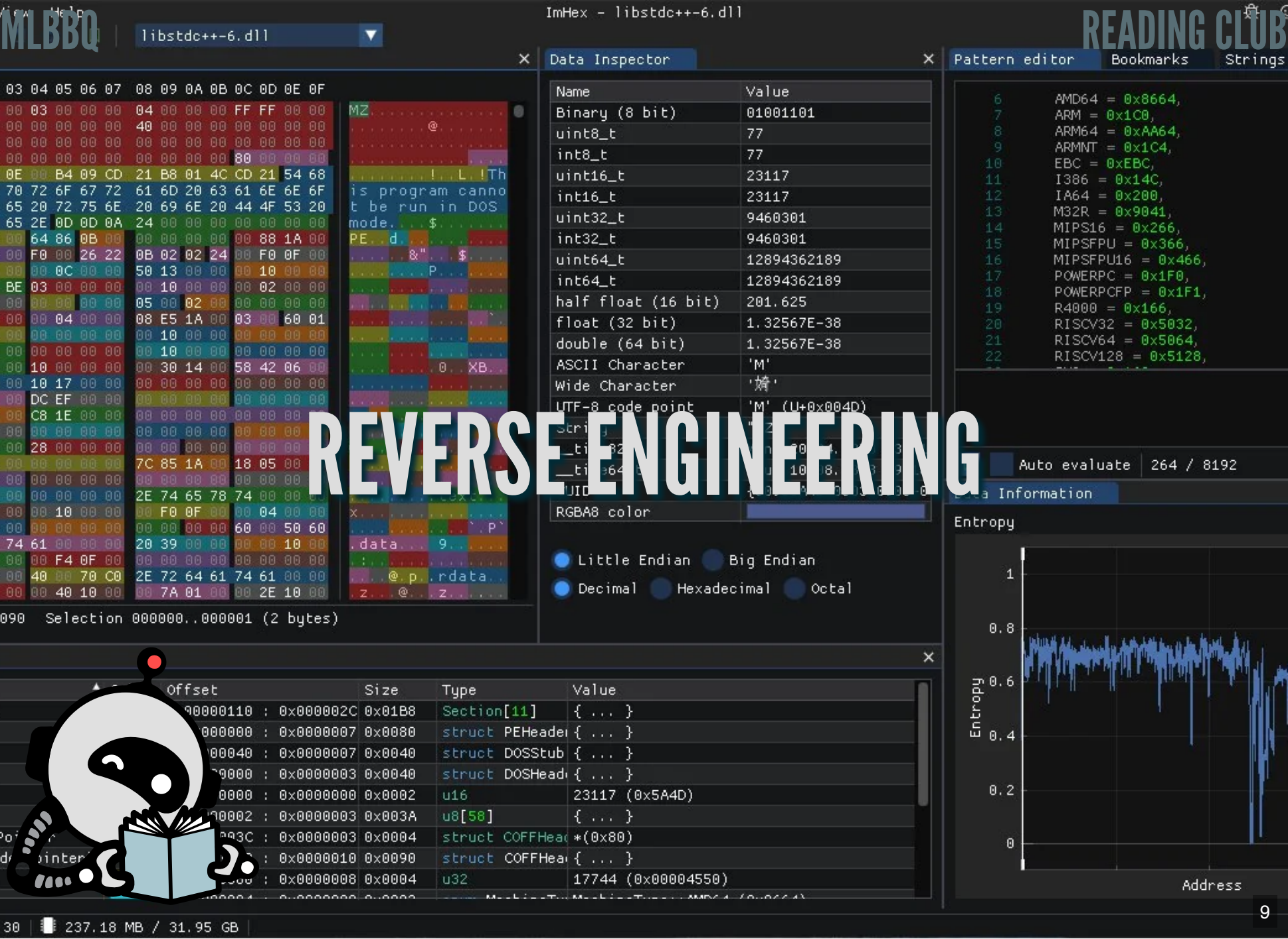


HOMEWORK PROBLEM

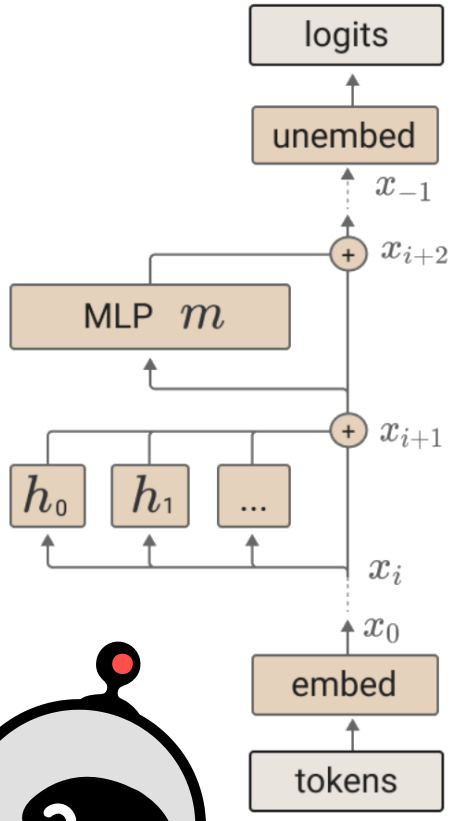


"Geometrically Guided Saliency Maps" Md Mahfuzur Rahman, Noah Lewis, Sergey Plis ICLR 2022 Workshop PAIR2Struct

REVERSE ENGINEERING



TRANSFORMER BLOCK



The final logits are produced by applying the unembedding.

$$T(t) = W_U x_{-1}$$

An MLP layer, m, is run and added to the residual stream.

$$x_{i+2} = x_{i+1} + m(x_{i+1})$$

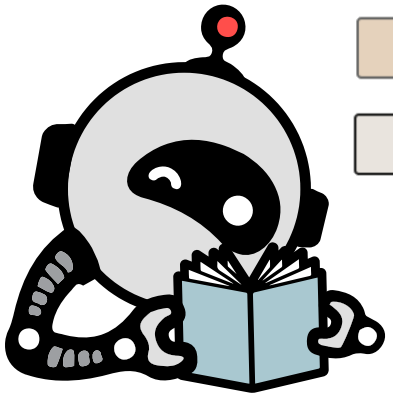
Each attention head, h, is run and added to the residual stream.

$$x_{i+1} = x_i + \sum_{h \in H_i} h(x_i)$$

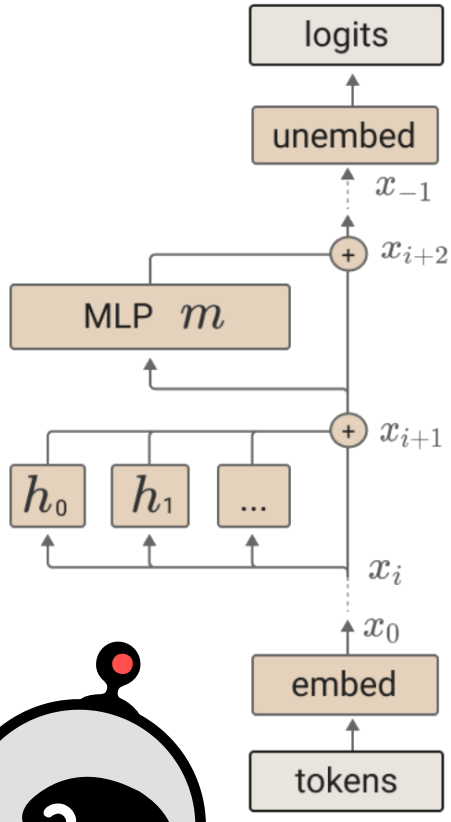
One residual block

Token embedding.

$$x_0 = W_E t$$



TRANSFORMER BLOCK



The final logits are produced by applying the unembedding.

$$T(t) = W_U x_{-1}$$

An MLP layer, m, is run and added to the residual stream.

$$x_{i+2} = x_{i+1} + m(x_{i+1})$$

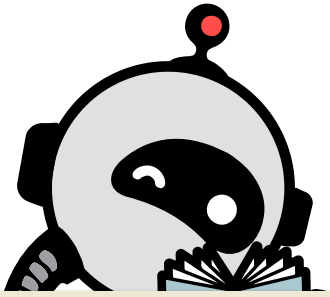
Each attention head, h, is run and added to the residual stream.

$$x_{i+1} = x_i + \sum_{h \in H_i} h(x_i)$$

One residual block

Token embedding.

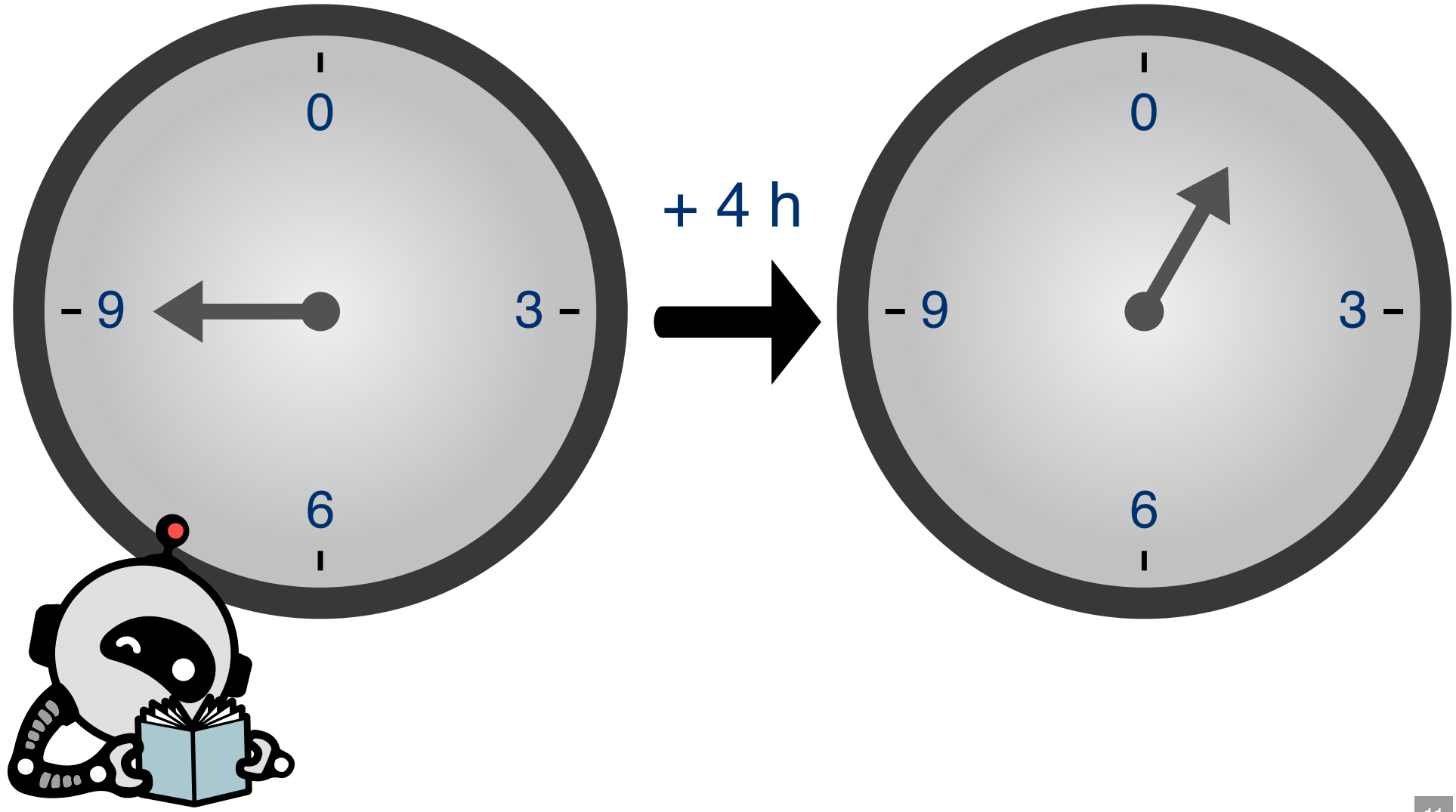
$$x_0 = W_E t$$



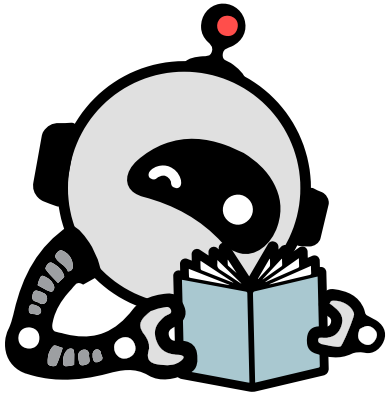
Knowing the parameter values tell us what they are doing conceptually

"Geometrically Guided Saliency Maps" Md Mahfuzur Rahman, Noah Lewis, Sergey Plis ICLR 2022 Workshop PAIR2Struct

THE TRANSFORMER WAS TRAINED ON MODULAR ADDITION



GROKKING



Grok /ˈgrɒk/ is a neologism coined by American writer Robert A. Heinlein for his 1961 science fiction novel Stranger in a Strange Land. While the Oxford English Dictionary summarizes the meaning of grok as "to understand intuitively or by empathy, to establish rapport with" and "to empathize or communicate sympathetically (with); also, to experience enjoyment",[1] Heinlein's concept is far more nuanced, with critic Istvan Csicsery-Ronay Jr. observing that "the book's major theme can be seen as an extended definition of the term."[2] The concept of grok garnered significant critical scrutiny in the years after the book's initial publication. The term and aspects of the underlying concept have become part of communities such as computer science.

grok /grŏk/

transitive verb

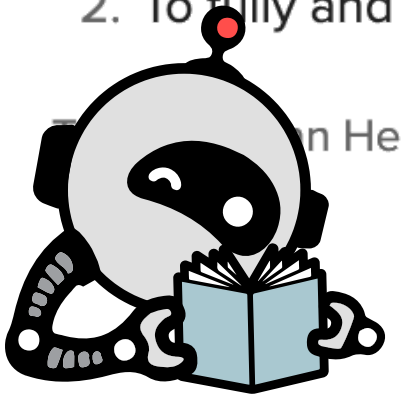
1. To understand profoundly through intuition or empathy.

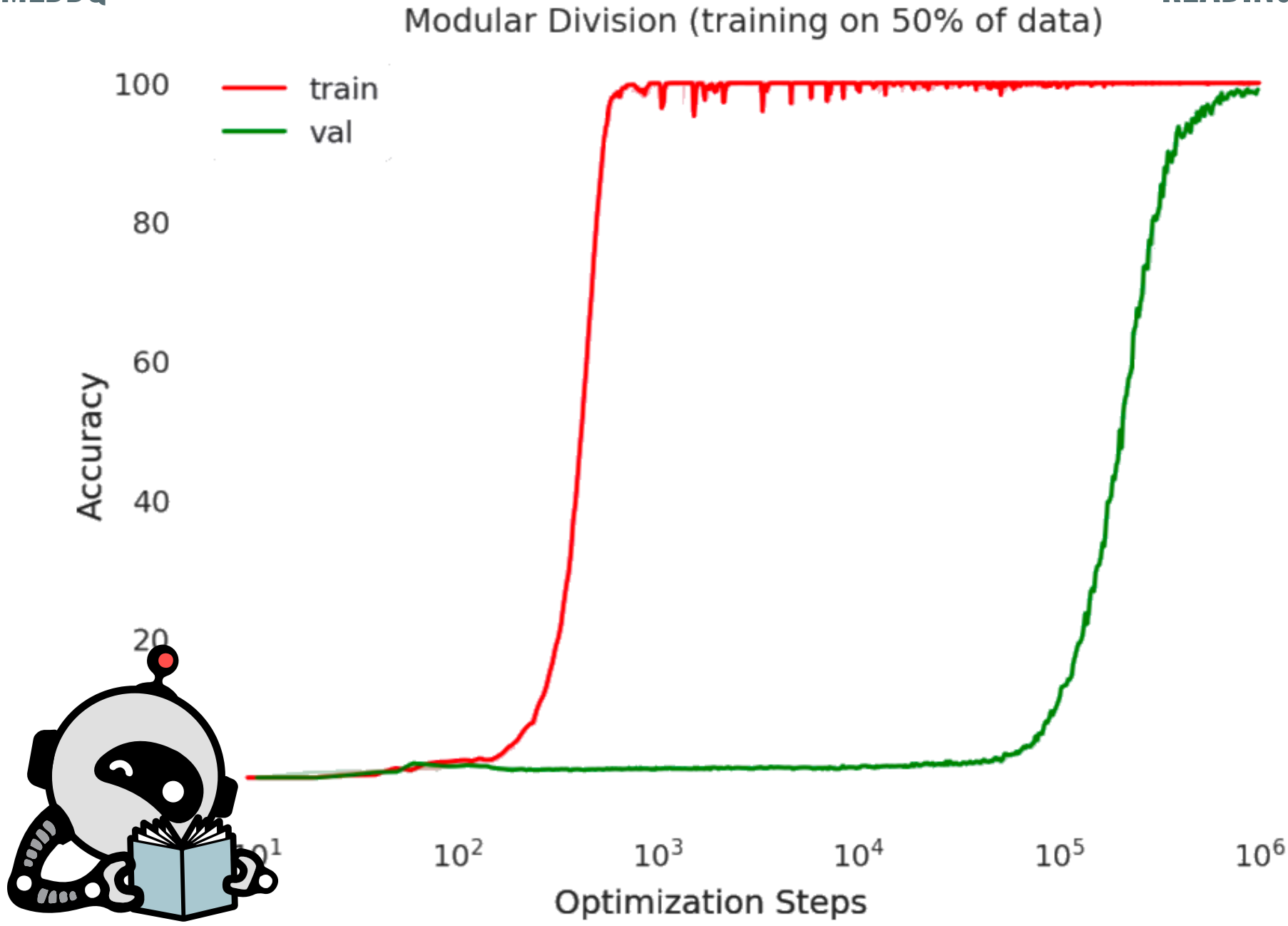
verb

1. To have or to have acquired an **intuitive understanding** of; to **know** (something) without having to **think** (such as knowing the number of objects in a collection without needing to count them: see **subitize**).

2. To fully and completely understand something in all its details and intricacies.

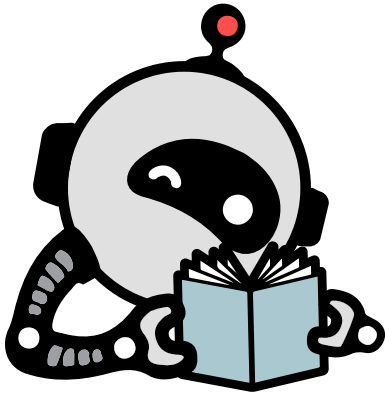
n Heritage® Dictionary of the English Language, 5th Edition • More at Wordnik





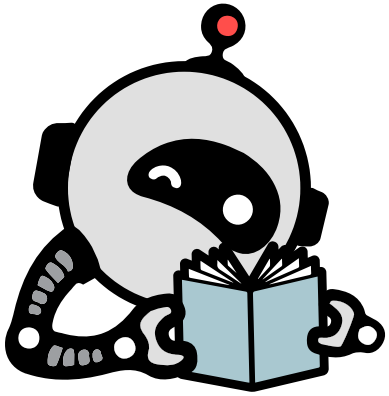
"Geometrically Guided Saliency Maps" Md Mahfuzur Rahman, Noah Lewis, Sergey Plis ICLR 2022 Workshop PAIR2Struct

LEARNING CURVES



"Geometrically Guided Saliency Maps" Md Mahfuzur Rahman, Noah Lewis, Sergey Plis ICLR 2022 Workshop PAIR2Struct

PROGRESS MEASURES



"Geometrically Guided Saliency Maps" Md Mahfuzur Rahman, Noah Lewis, Sergey Plis ICLR 2022 Workshop PAIR2Struct