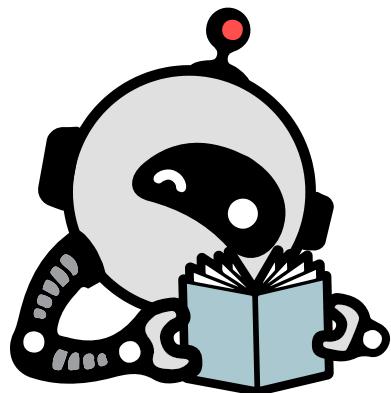


# PROGRESS MEASURES FOR GROKKING VIA MECHANISTIC INTERPRETABILITY

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

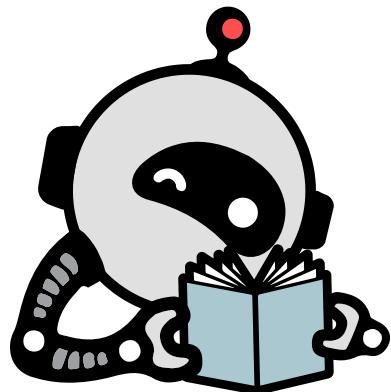


# OUTLINE



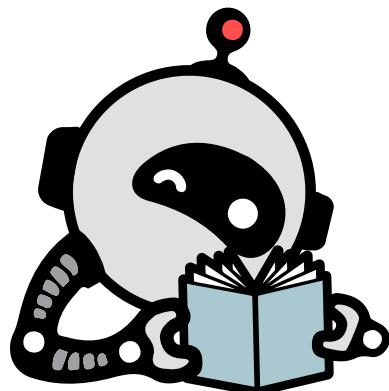
# OUTLINE

- Mechanistic Interpretability



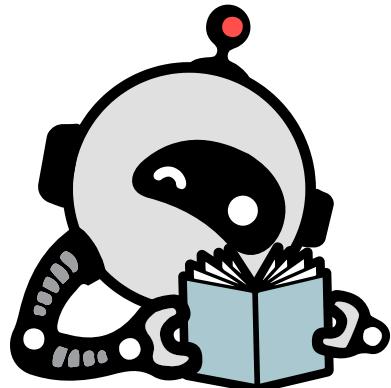
# OUTLINE

- Mechanistic Interpretability
- grok /grōk/

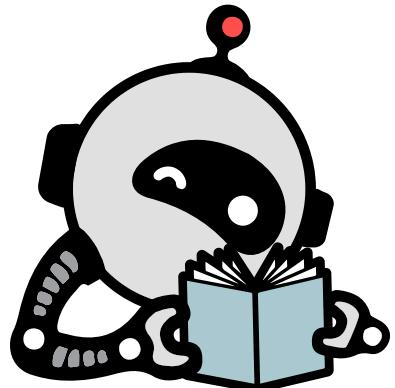


# OUTLINE

- Mechanistic Interpretability
- grok /grōk/
- Progress Measures



# MECHANISTIC INTERPRETABILITY



# TWO TYPES OF EXPLAINABLE AI

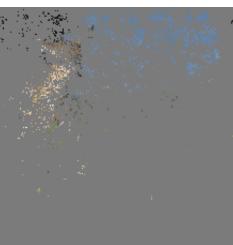
## EXPLAIN THE DATA

## INTERPRET THE MODEL

Input



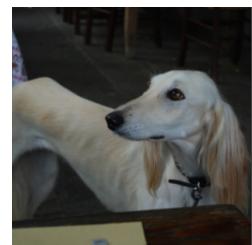
Gradients



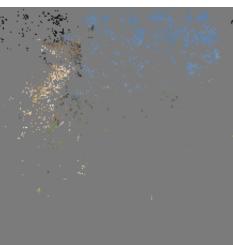
Integrated Gradients



our WIP method



Gradients



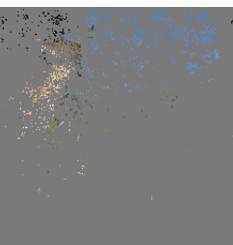
Integrated Gradients



our WIP method



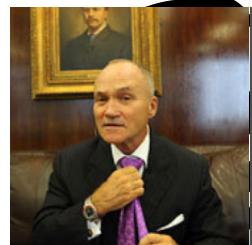
Gradients



Integrated Gradients



our WIP method



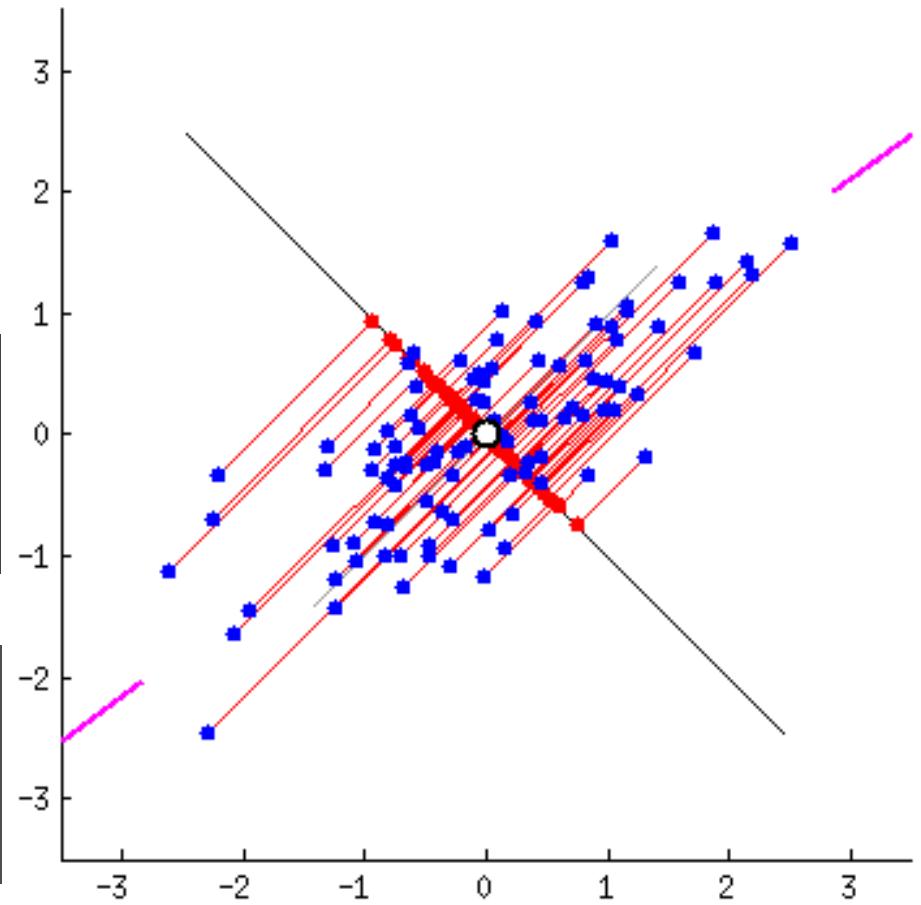
Gradients



Integrated Gradients



our WIP method



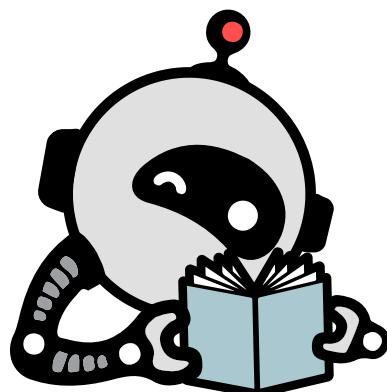
# HOMEWORK PROBLEM

*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



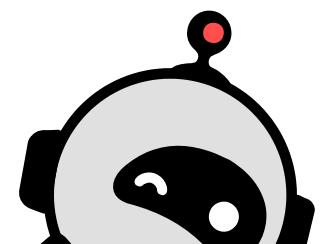
# HOMEWORK PROBLEM

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

$$100111 + 110010 = 1011001$$

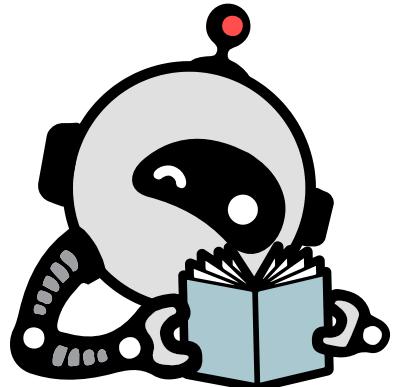
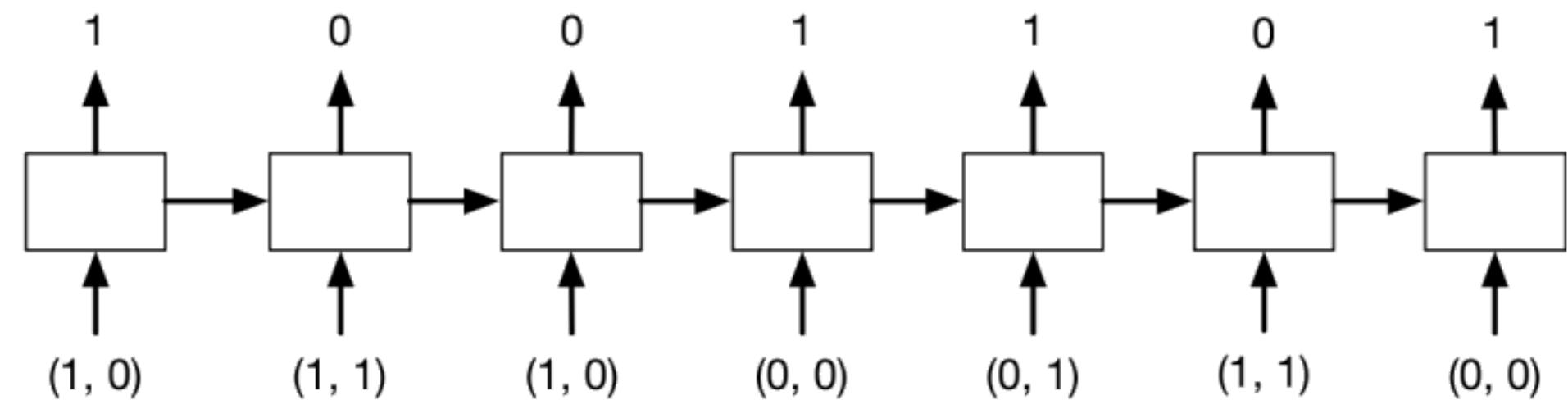
would be represented as :

- 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

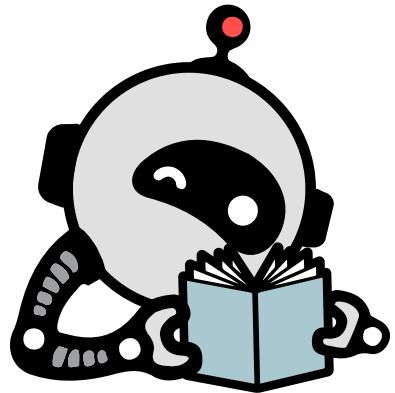
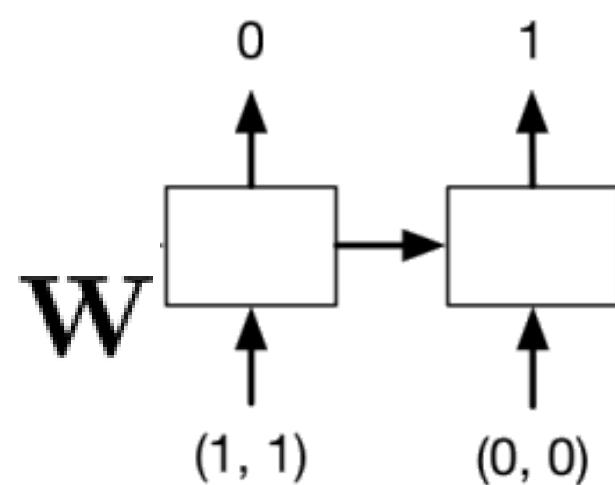
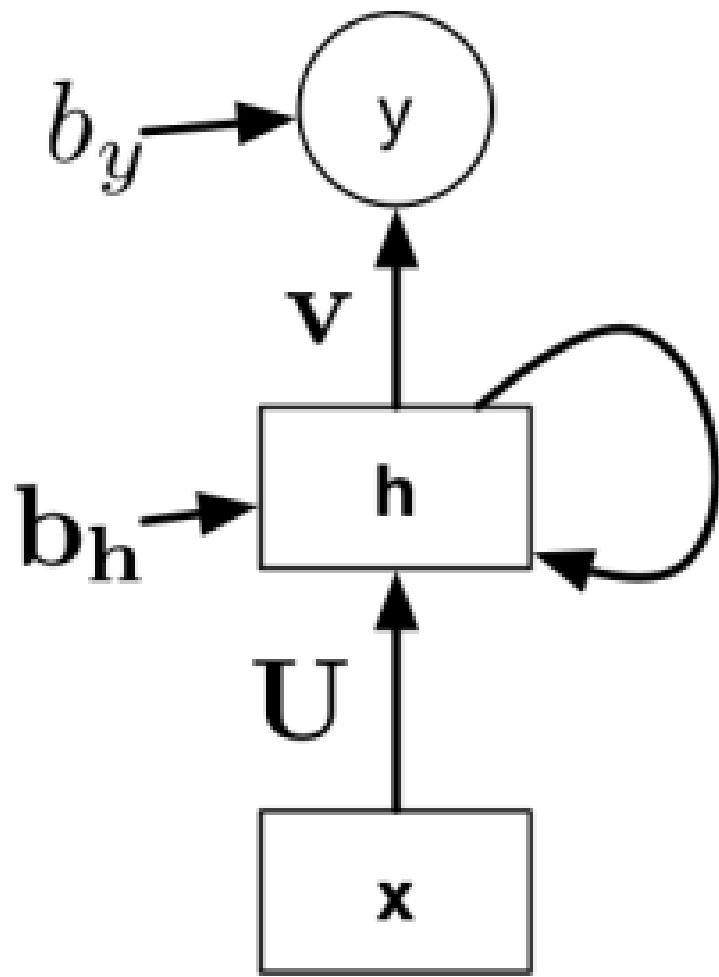
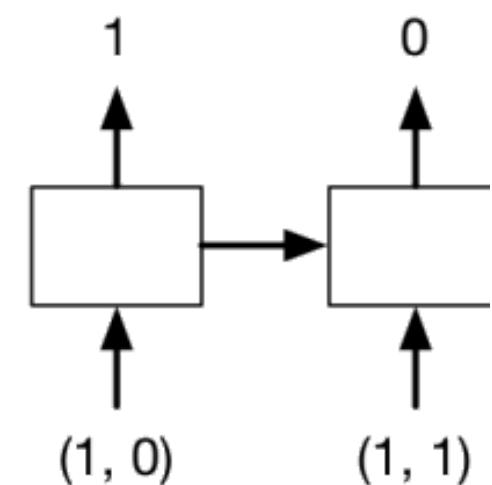


*Set the parameter values manually - no training!*

# HOMEWORK PROBLEM



# HOMEWORK PROBLEM



File Help

libstdc++-6.dll

ImHex - libstdc++-6.dll

**READING CLUB**

Data Inspector

Name	Value
Binary (8 bit)	01001101
uint8_t	77
int8_t	77
uint16_t	23117
int16_t	23117
uint32_t	9460301
int32_t	9460301
uint64_t	12894362189
int64_t	12894362189
half float (16 bit)	201.625
float (32 bit)	1.32567E-38
double (64 bit)	1.32567E-38
ASCII Character	'M'
Wide Character	'M'
UTF-8 code point	'M' (U+0x004D)
String	"Z"
_ati_32	"Z"
_ati_64	"Z"
UUID	{00000000-0000-0000-0000-000000000000}
RGBAB color	

Auto evaluate 264 / 8192

Entropy

REVERSE ENGINEERING

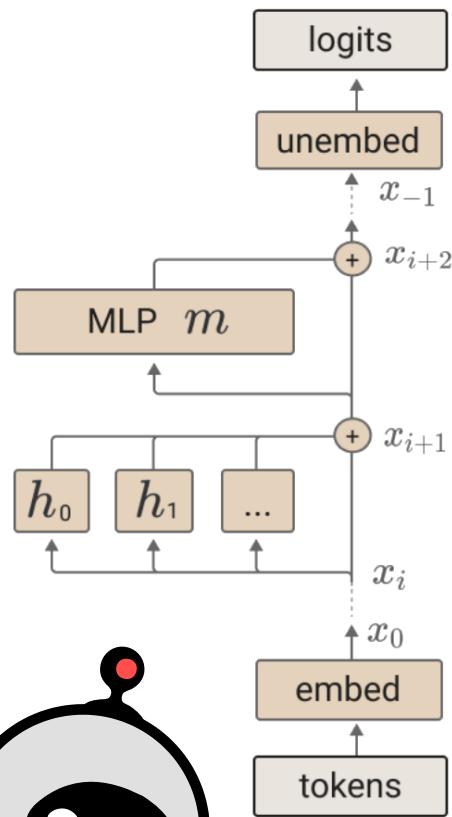
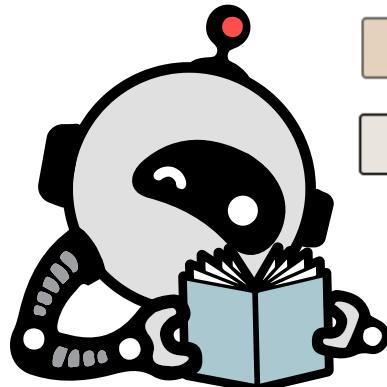
Little Endian   Big Endian  
 Decimal   Hexadecimal   Octal



Selection 000000..0000001 (2 bytes)

Offset	Size	Type	Value
000000110 : 0x0000002C	0x01B8	Section[11]	{ ... }
000000 : 0x00000007	0x0080	struct PEHeader	{ ... }
000040 : 0x00000007	0x0040	struct DOSStub	{ ... }
0000 : 0x00000003	0x0040	struct DOSHeader	{ ... }
0000 : 0x00000000	0x0002	u16	23117 (0x5A4D)
00002 : 0x00000003	0x003A	u8[58]	{ ... }
0003C : 0x00000003	0x0004	struct COFFHeader	* (0x80)
00040 : 0x00000010	0x0090	struct COFFHeader	{ ... }
00040 : 0x00000008	0x0004	u32	17744 (0x00004550)
0000001 : 0x00000000	0x0000		

# 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.

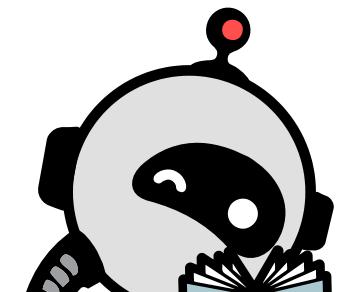
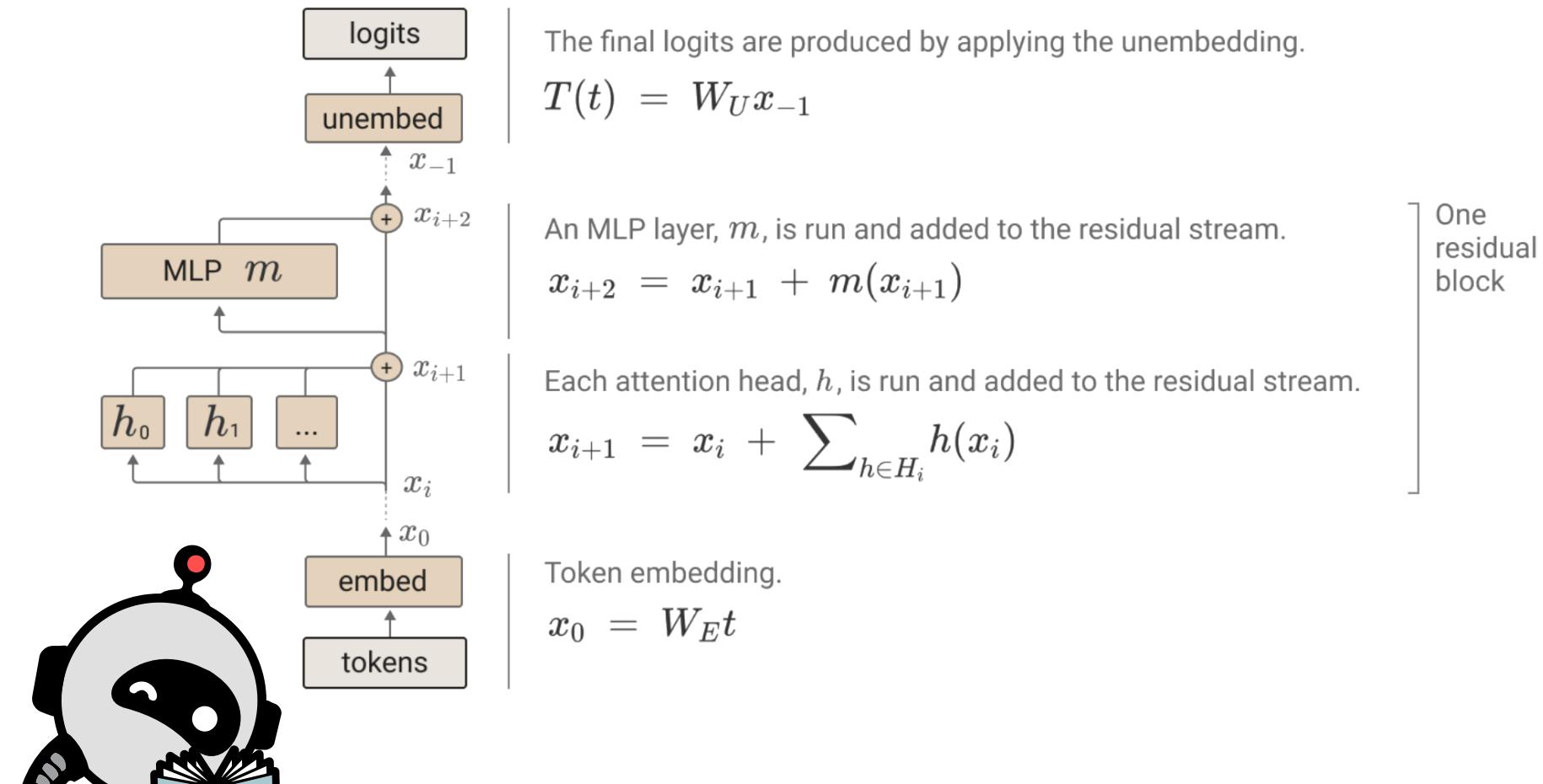
$$x_{i+1} = x_i + \sum_{h \in H_i} h(x_i)$$

One residual block

Token embedding.

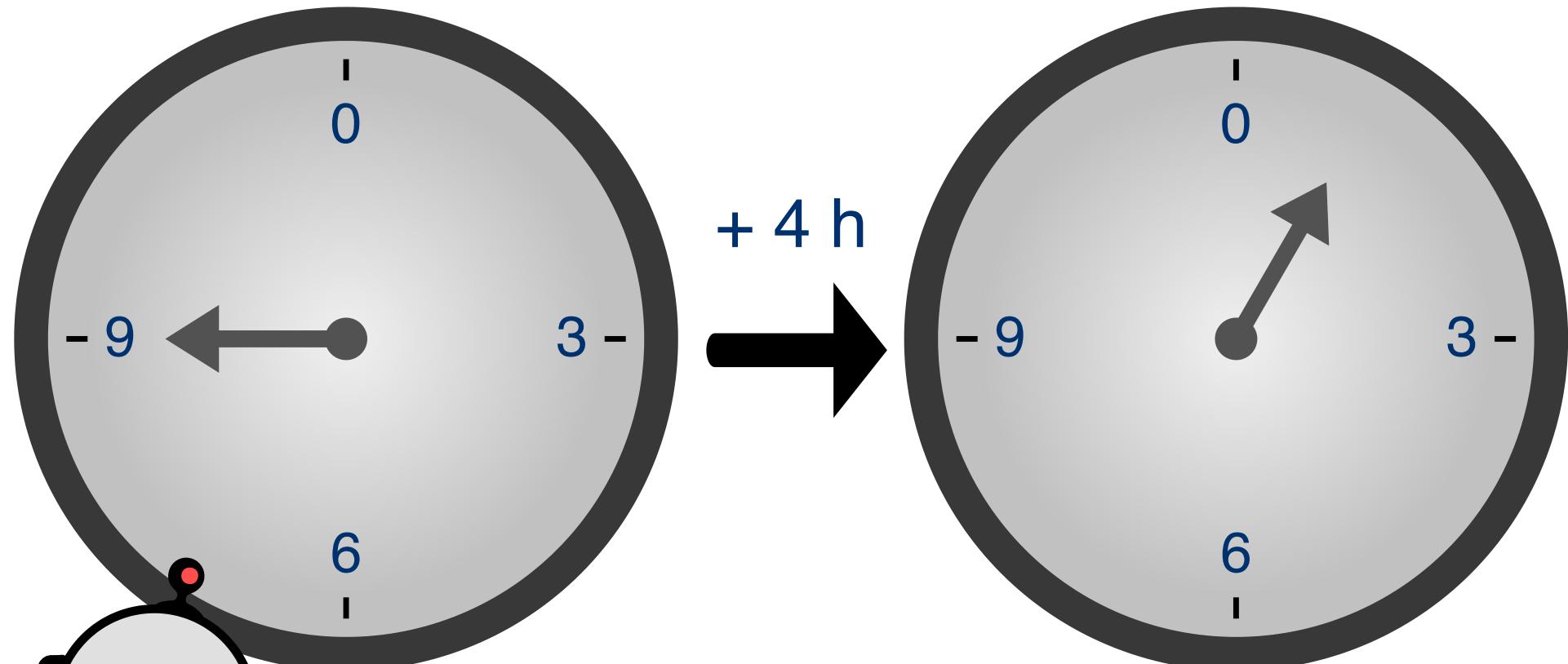
$$x_0 = W_E t$$

# TRANSFORMER BLOCK

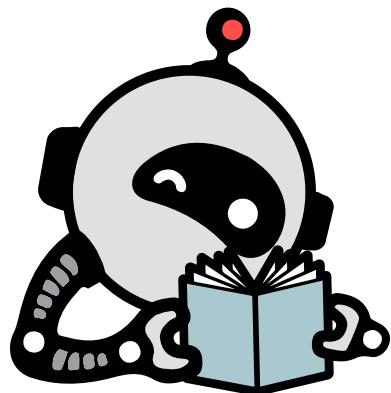


*Knowing the parameter values tell us what they are doing conceptually*

# THE TRANSFORMER WAS TRAINED ON MODULAR ADDITION



# GROKKING



Grok /'grɒk/ 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 /grōk/

**transitive verb**

1. To understand profoundly through intuition or empathy.

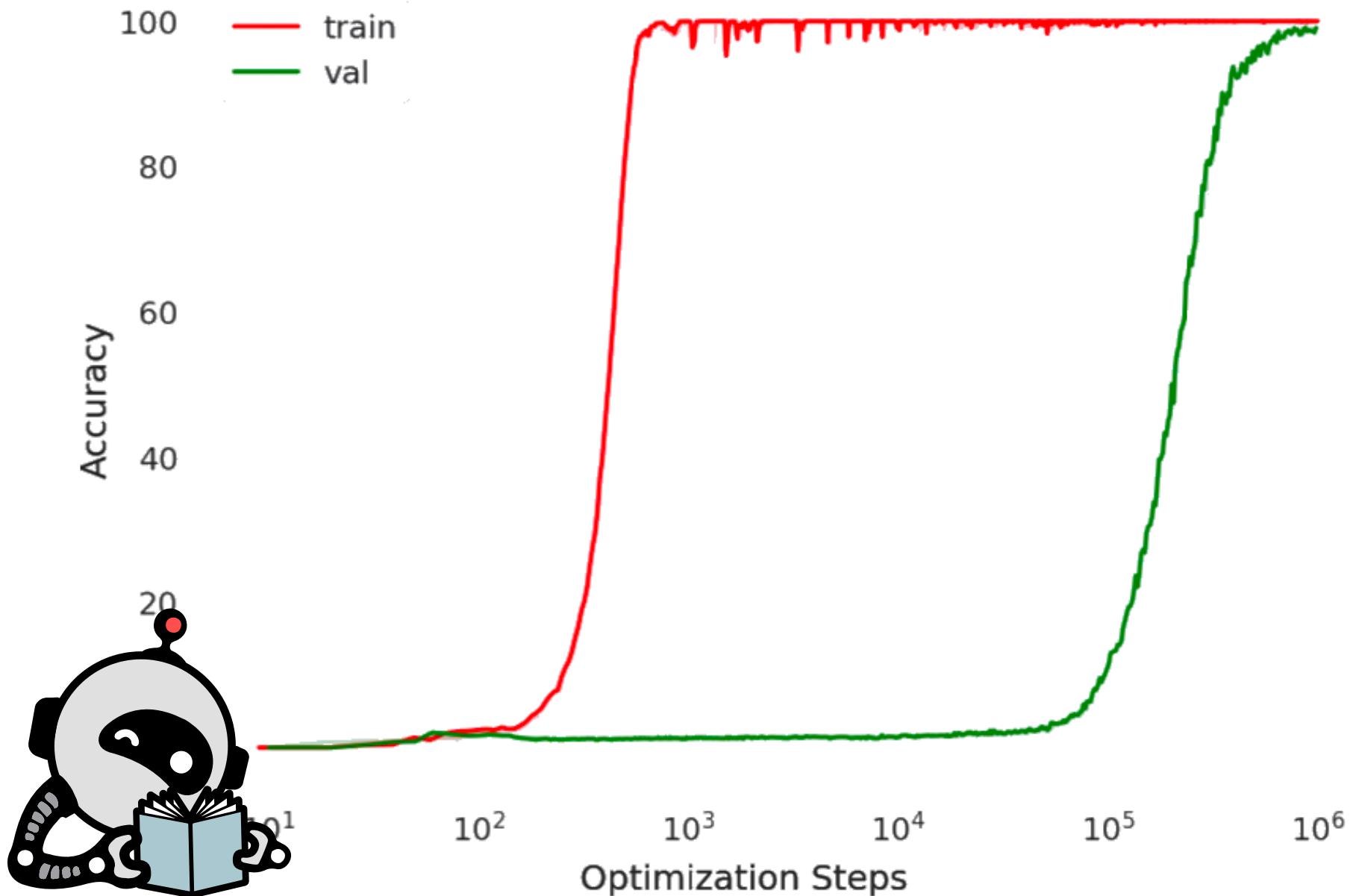
**verb**

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.

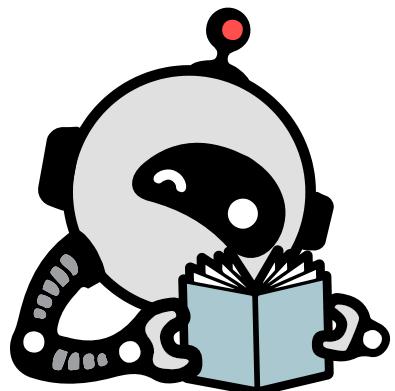


From the Oxford English Dictionary Heritage® Dictionary of the English Language, 5th Edition • More at [Wordnik](#)

## Modular Division (training on 50% of data)



# LEARNING CURVES



# PROGRESS MEASURES

