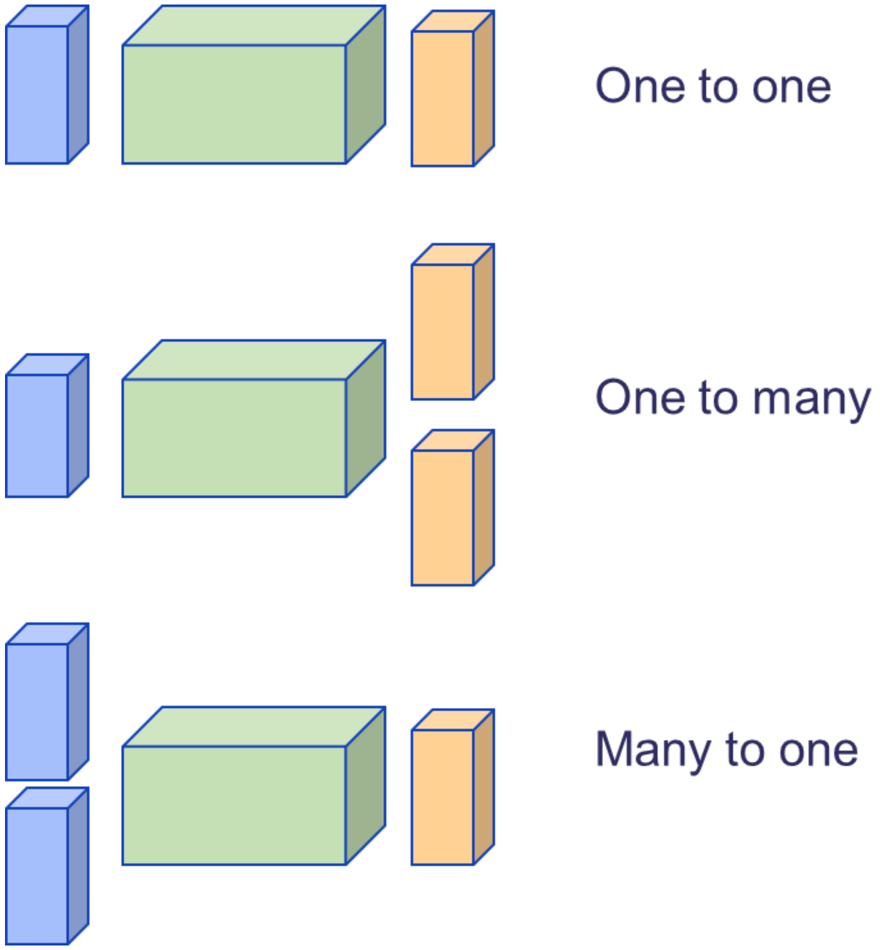
SEQUENCES AND LANGUAGE MODELS

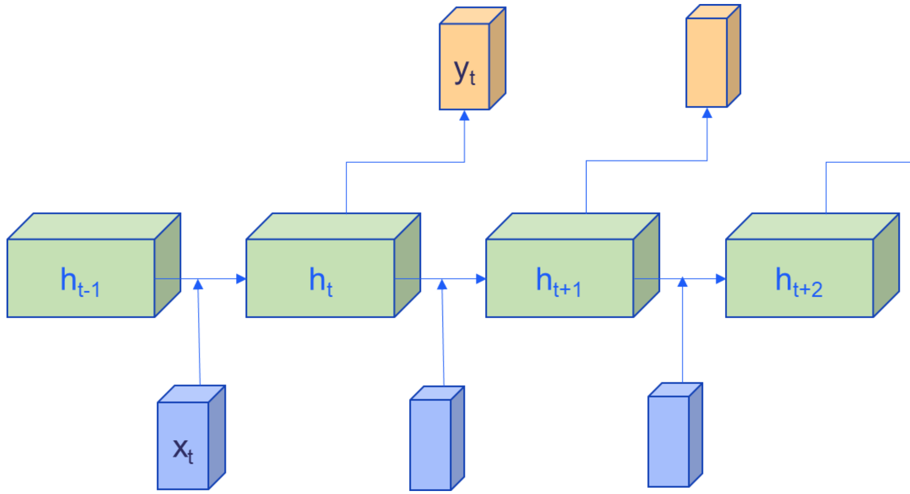
Keith Butler

SEQUENCE DATA



MLPs and CNNs struggle with sequence data. There is no fixed length of input, the data can be many to one.

RECURRENT NEURAL NETWORKS (RNNS)



$$h_t = f_W(W_{hh}h_{t-1}, W_{xh}x_t)$$

ISSUES WITH RNNS

RNNs have very short term memory – they lose context quickly

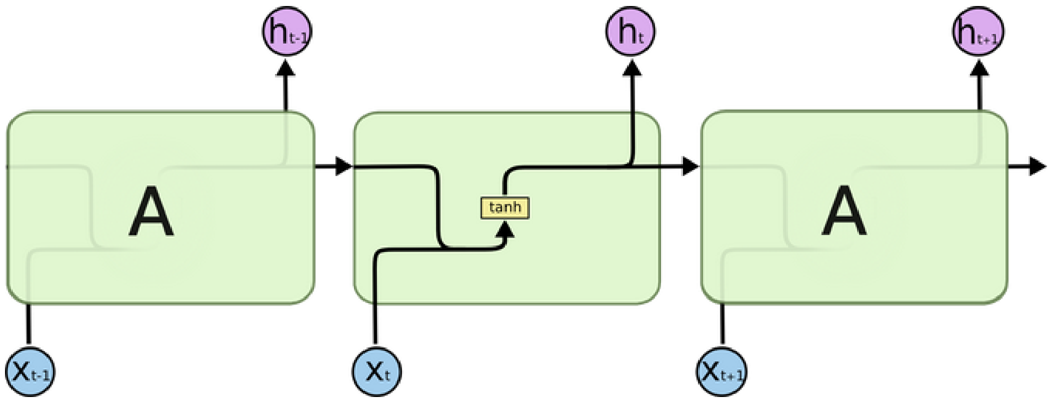
The clouds are in the _________. ✓

I was born in France. At the age of 16 I moved country. I have lived here since I was 21. Nonetheless, I still speak fluent ________. ✗

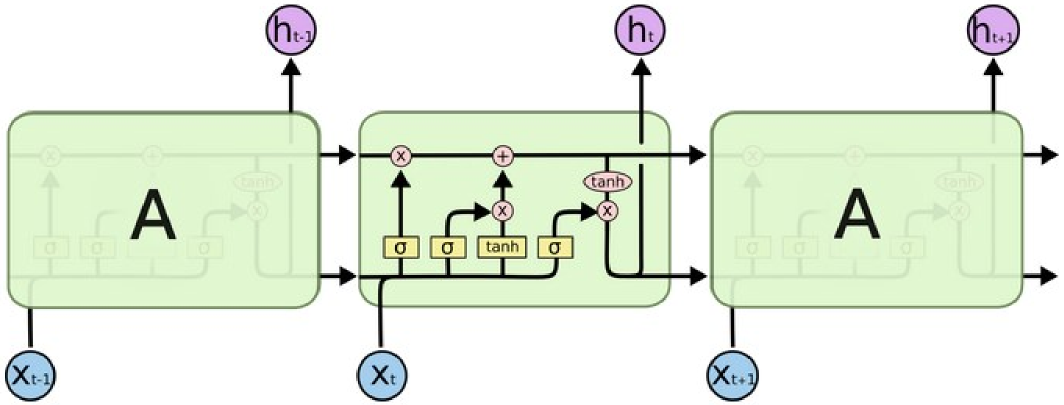
https://colah.github.io/posts/2015-08-Understanding-LSTMs/

INTRODUCING MEMORY

RNN



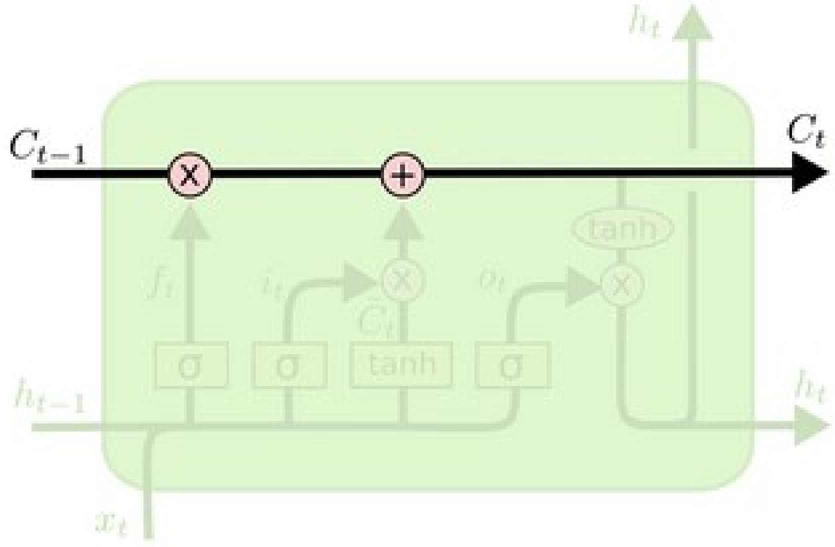
LSTM



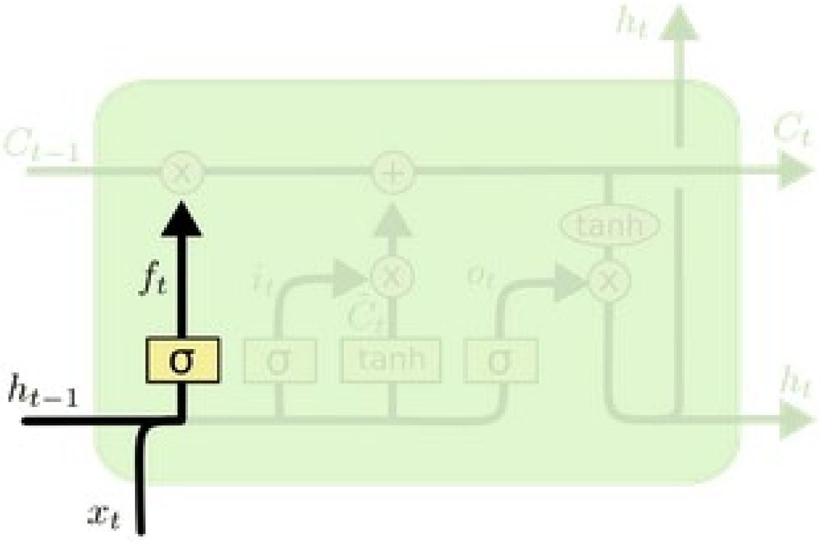
Long short term memory (LSTM) networks introduce extra memory features compared to a standard RNN

LSTM – THE MEMORY STATE

A single channel that runs all the way along the sequence structure
Only has some minor interactions with the rest of the network

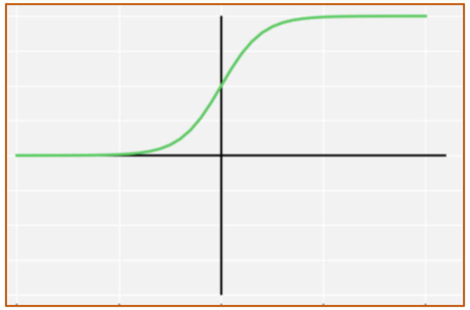


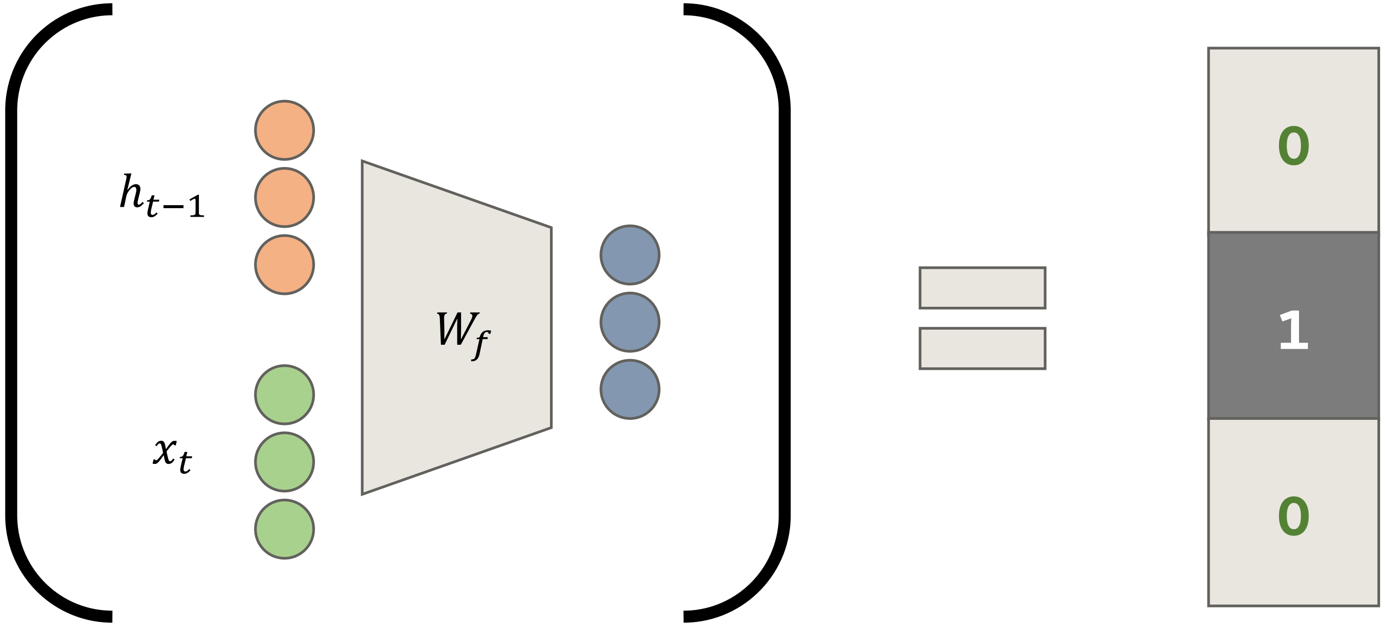
LSTM – FORGETTING



$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] + b_f\right)$$

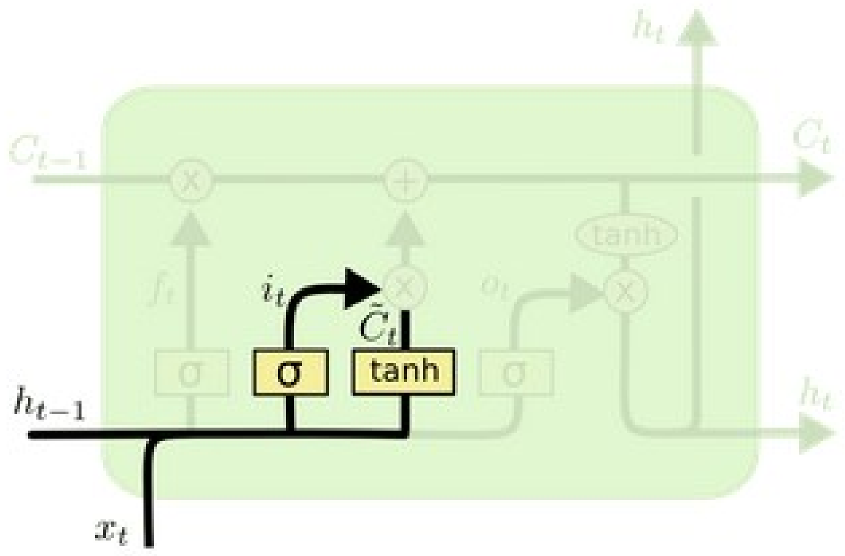
FORGETTING





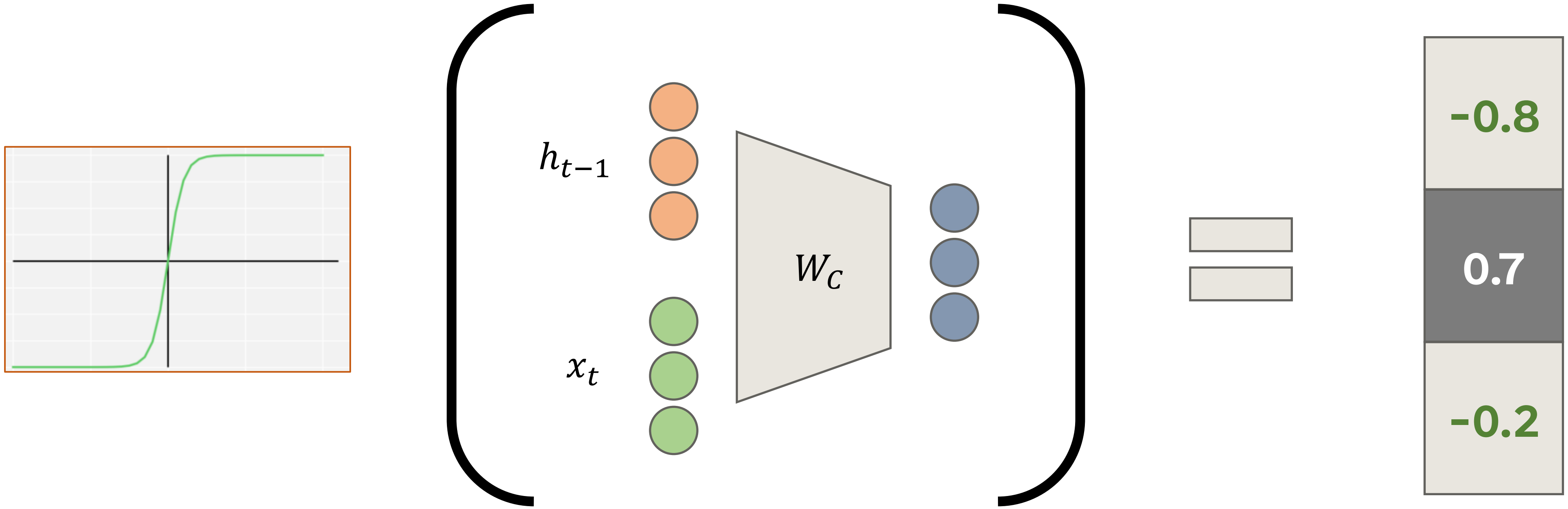
$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] + b_f\right)$$

LSTMS - ADDING



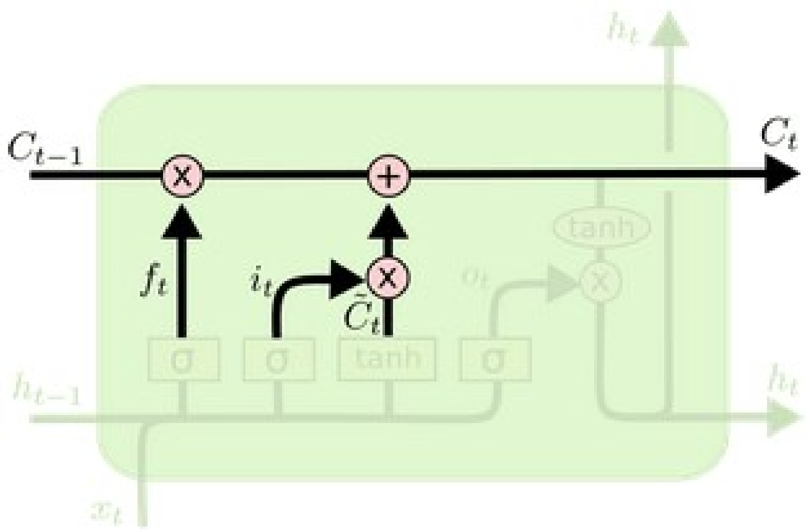
$$i_t = \sigma\left(W_i \cdot [h_{t-1}, x_t] + b_i\right)$$
$$\tilde{C}_t = tanh\left(W_C \cdot [h_{t-1}, x_t] + b_C\right)$$

ADDING



$$\tilde{C}_t = tanh\left(W_C \cdot [h_{t-1}, x_t] + b_C\right)$$

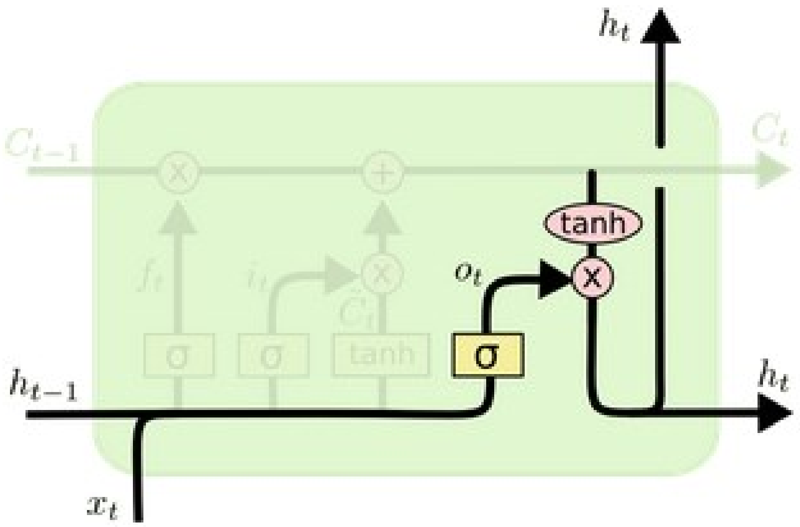
LSTM – UPDATING THE MEMORY



$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$$

LSTM – GENERATE OUTPUT

student



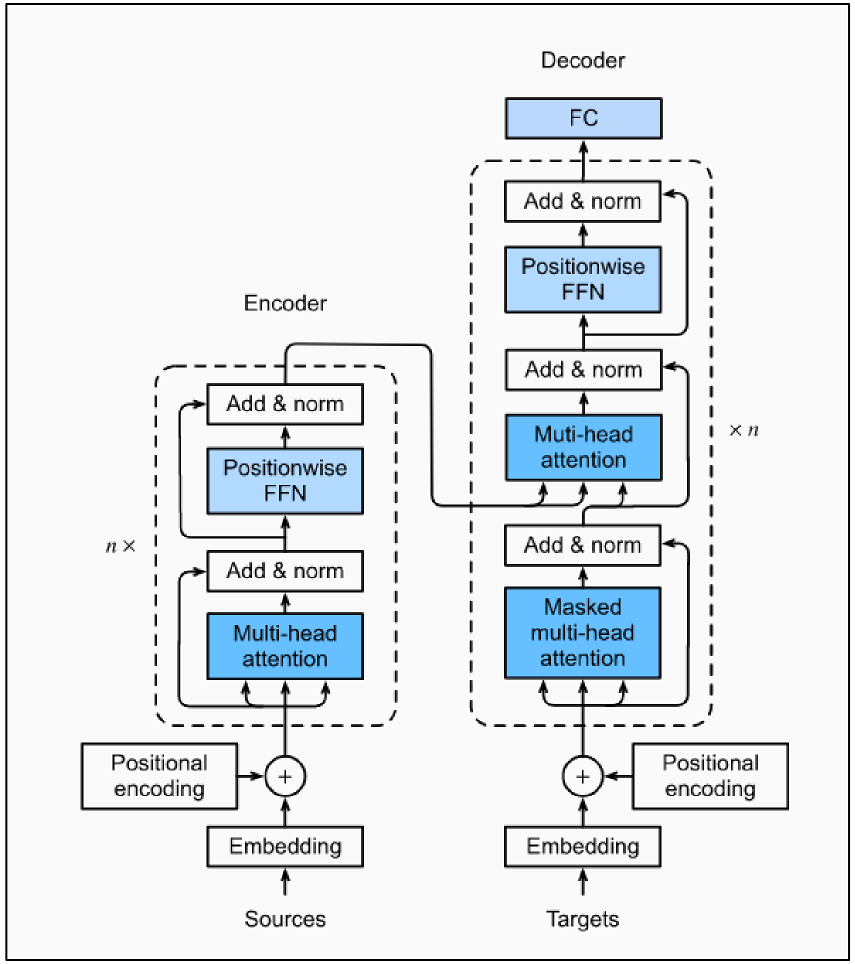
Je suis etudiant

$$o_t = \sigma\left(W_o\left[h_{t-1}, x_t\right] + b_o\right)$$

$$h_t = o_t * \tanh(C_t)$$

TRANSFORMERS

student



Je suis etudiant I am a

THE ATTENTION MECHANISM

High attention

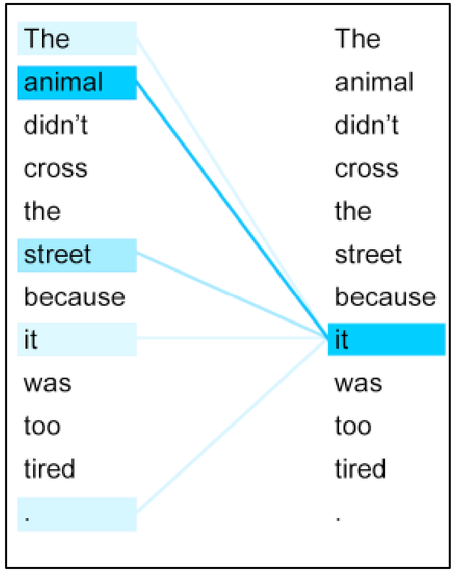
Low attention

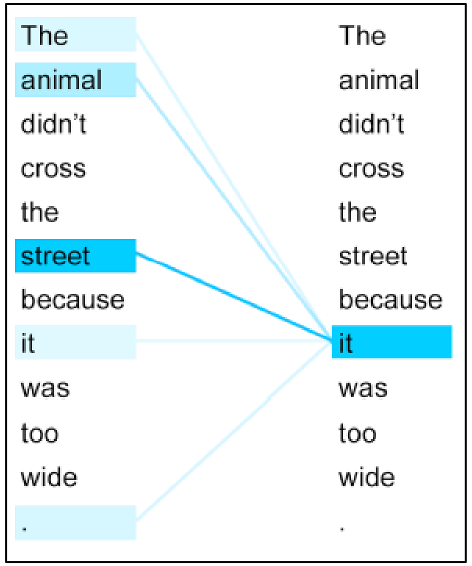
She is eating a green apple.

We use attention to “focus” on some part of interest in an input

SELF-ATTENTION

With self-attention, each token $\boldsymbol{t}_n$ can "attend to" all other tokens of the same sequence when computing this token's embedding $\boldsymbol{x}_n$





HOW SELF-ATTENTION WORKS

Attention Is All You Need

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d}}\right)V$$

Based of the concept of query, key and value vectors

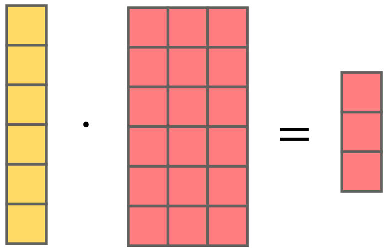
arXiv:1706.03762 (2017)

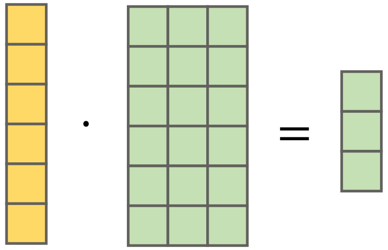
ATTENTION – THE INPUTS

Each token has a d-dimensional representation

Each token also has a query and key vector q-dimensional; q << d

W_K is a matrix of learnable weights





$\overrightarrow{x_i} \cdot W_Q = \overrightarrow{Q_i}$

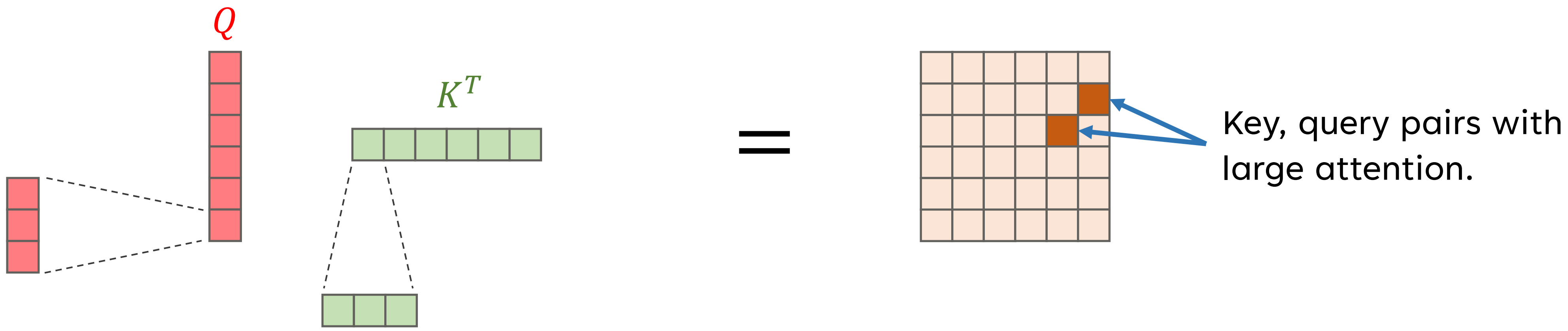
$\overrightarrow{x_i} \cdot W_K = \vec{K_i}$

In bed an old woman lies.

QUERY KEY MULTIPLICATION

$$softmax\left(\frac{QK^T}{\sqrt{d}}\right)$$

In bed an old woman lies.



Softmax normalises the columns; root of d makes it numerically stable

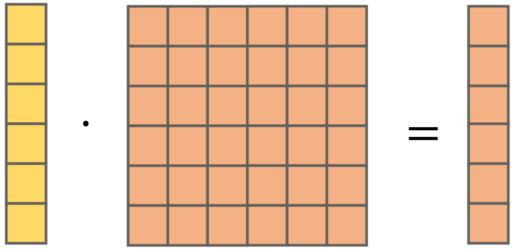
NOTE –in this case each cube in Q and K has 3 dimensions

THE VALUE MATRIX

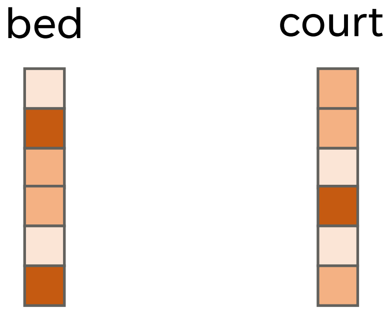
Tells you how a given token modifies another token

The resultant $\vec{V}$ gets added to the other vector

The extent of the addition is scaled by the QK^T product

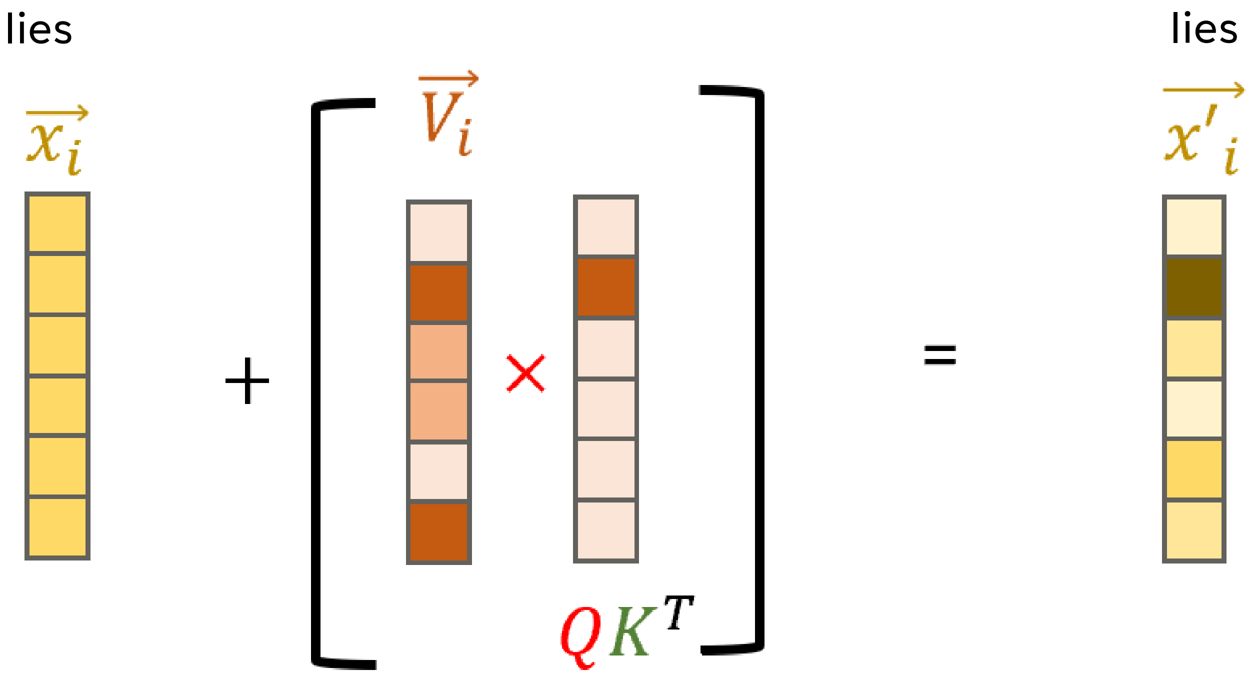


In ---- an old woman lies.



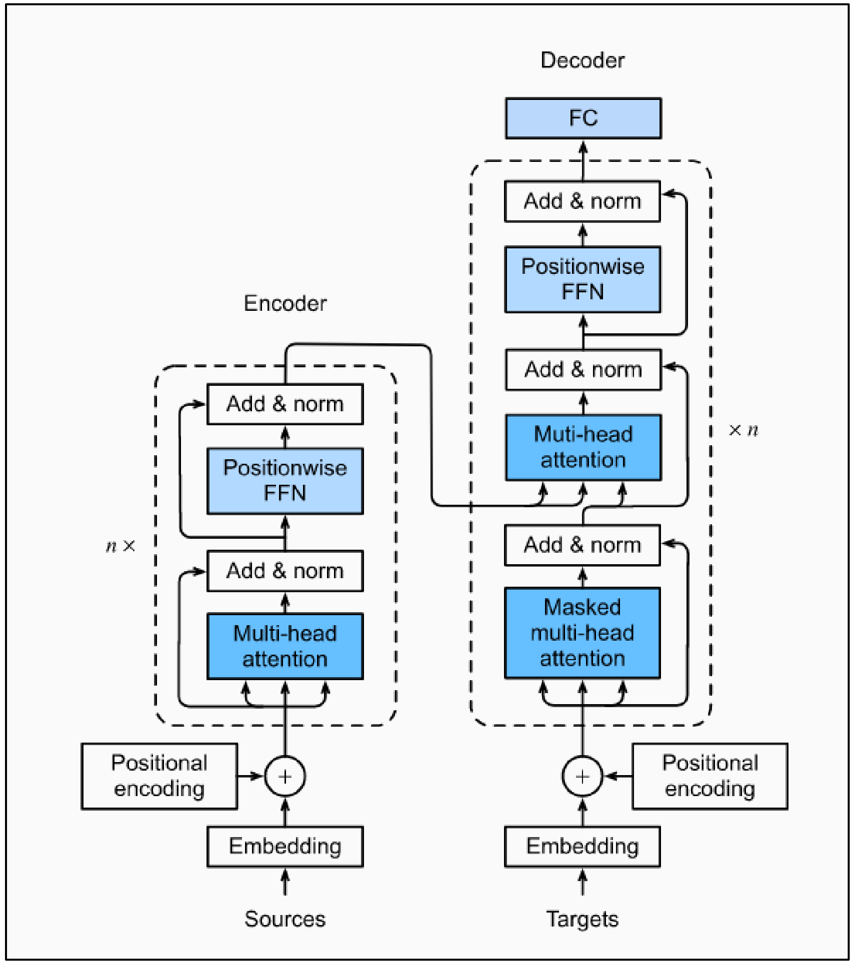
$$\vec{x_i} \cdot W_V = \vec{V_i}$$

ADDING THE VALUE TO THE EMBEDDING



The value is modified by the attention from the QK pair and added to the initial embedding

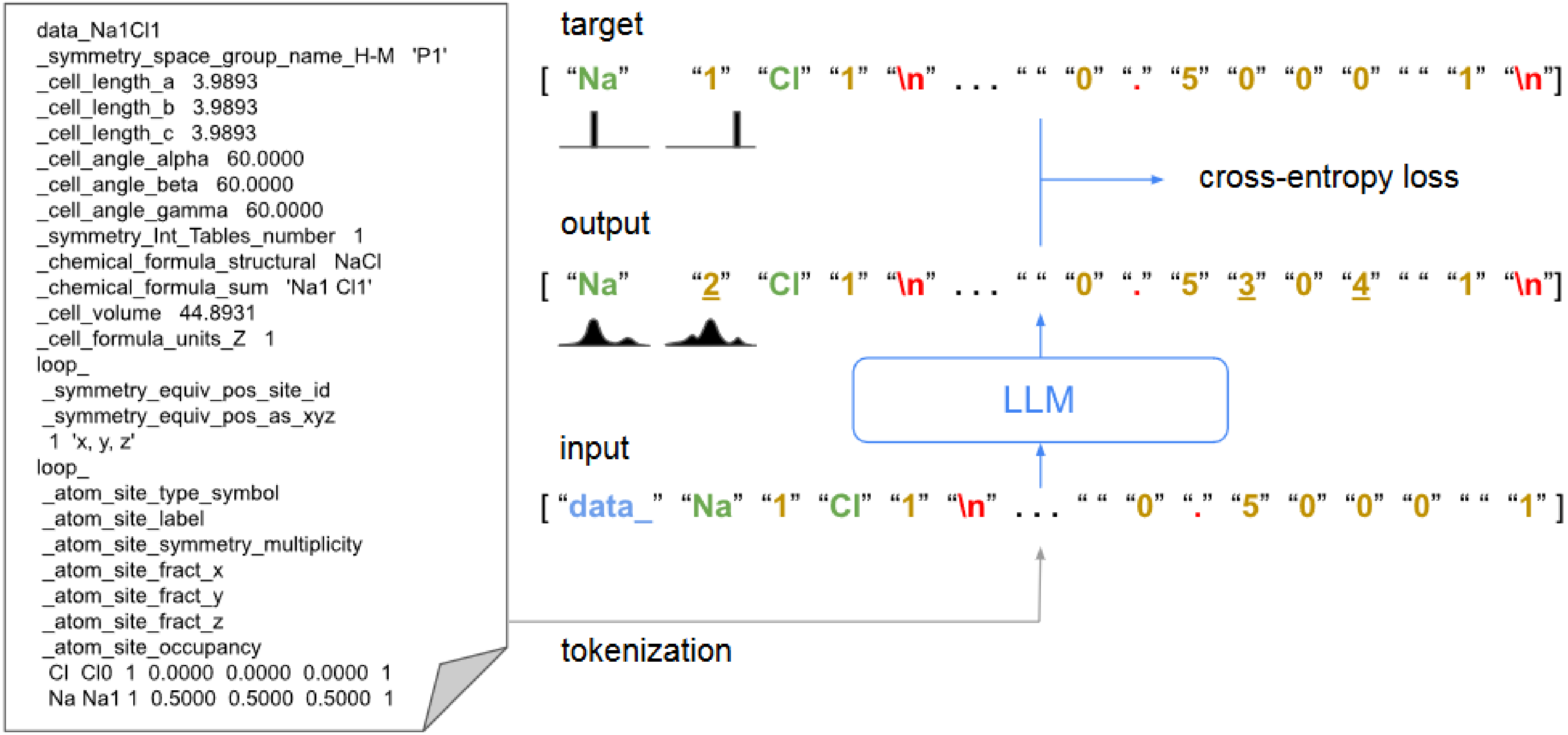
UPDATING THE EMBEDDINGS



$$E \rightarrow \boxed{softmax\left(\frac{QK^T}{\sqrt{d}}\right)V} \rightarrow E'$$

Each of these softmax matrix multiplications is a ‘head’

CRYSTALLM

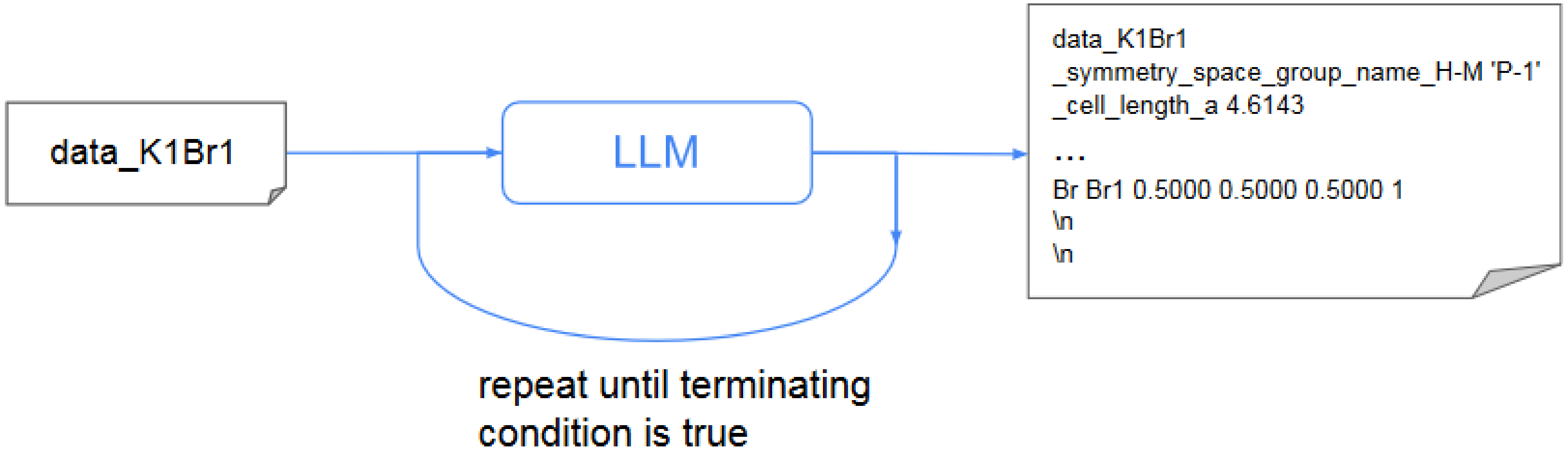


A decoder only transformer trained on cif files for materials structure generation

Nature Communications 15, 1 (2024)

AUTOREGRESSIVE GENERATION

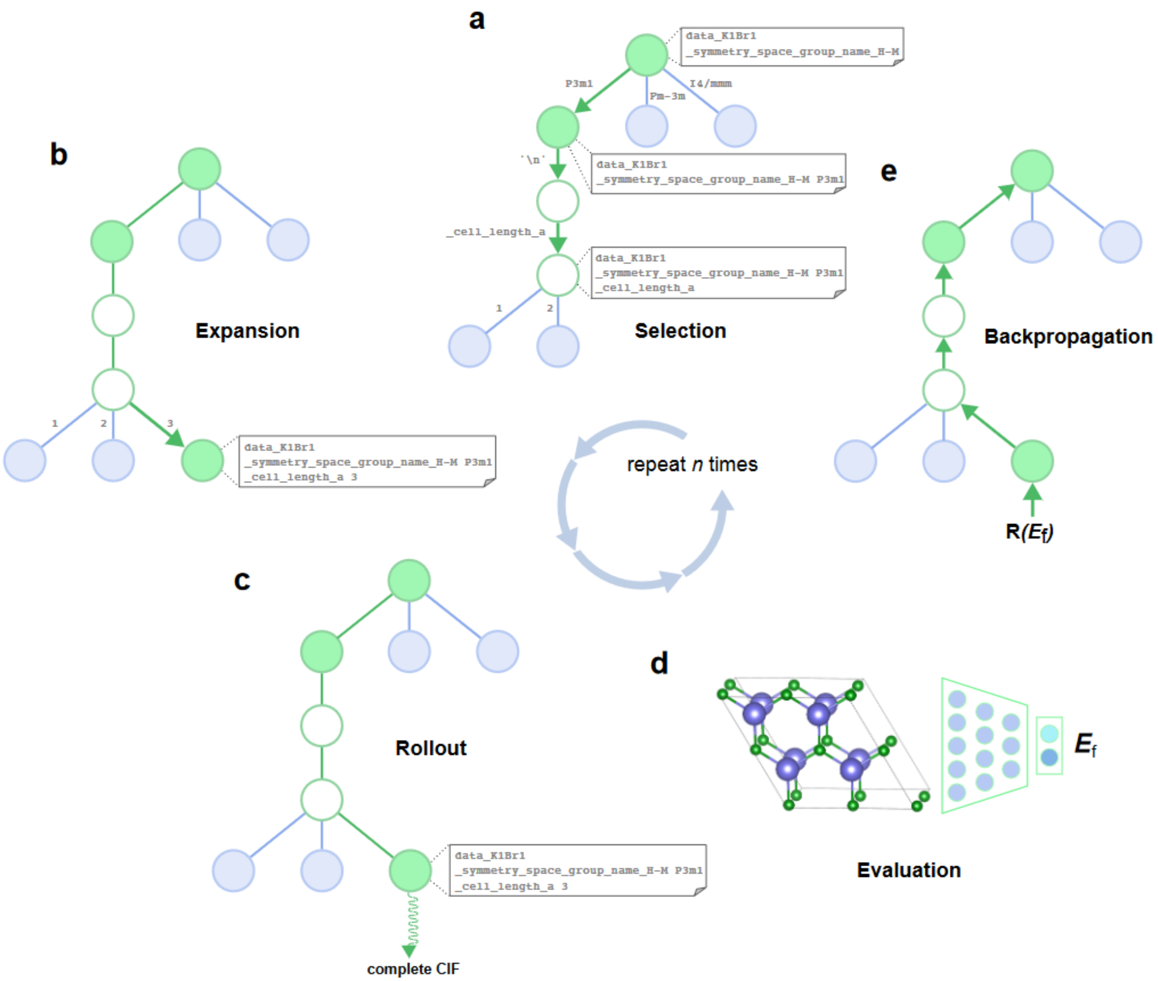
After training, CrystaLLM can be prompted with new text and can produce a predicted complete cif



Prompting is flexible so we can provide as little or as much information as we like

Nature Communications 15, 1 (2024)

MONTE CARLO TREE SEARCH FOR CONSISTENCY



Autoregressive generation is stochastic and can lead to non-ideal structures

MCTS is more expensive but uses an energy estimator to drive to low energy solutions

Nature Communications 15, 1 (2024)

CONCEPT CHECKLIST

Sequential data benefits from contextual awareness and memory

Recurrent networks are an early answer

Memory was improved with LSTMs

Transformers use attention to map across sequences

Autoregression can be applied to generate crystal structures

THANK YOU

mdi-group.github.com