An Alternate Formulation of Transformers

Large Language Models: Introduction and Recent Advances

ELL881 · AIL821

Tanmoy Chakraborty
Associate Professor, IIT Delhi
https://tanmoychak.com/

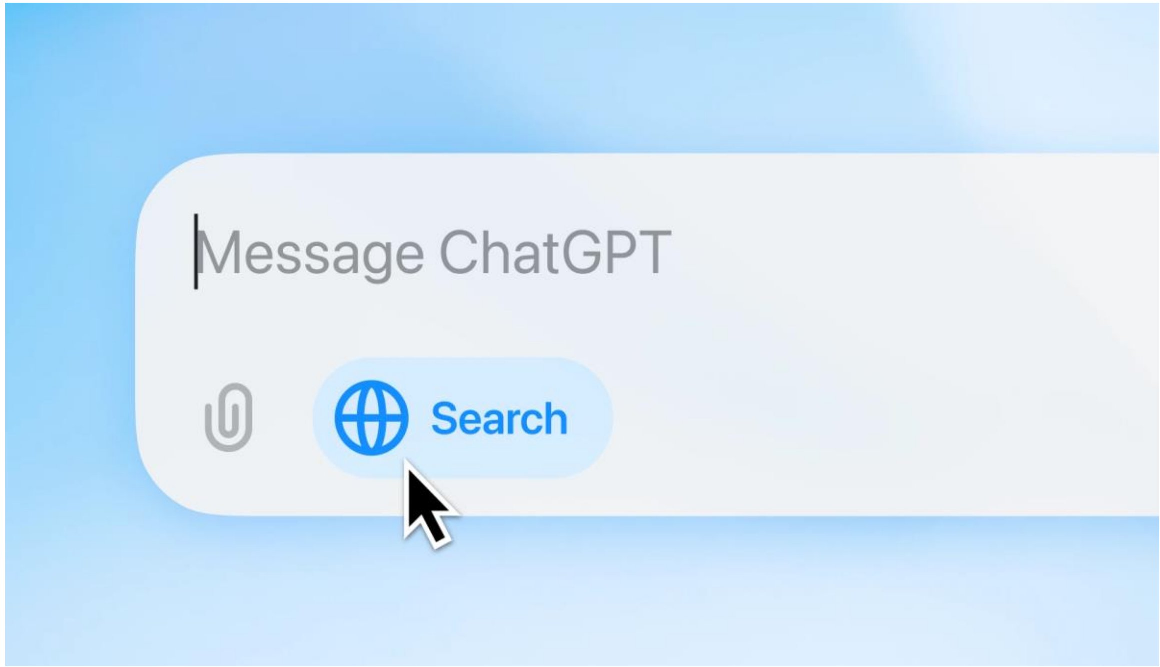
OpenAI introduces ChatGPT search !

Announced on
October 31, 2024

OpenAI Blog

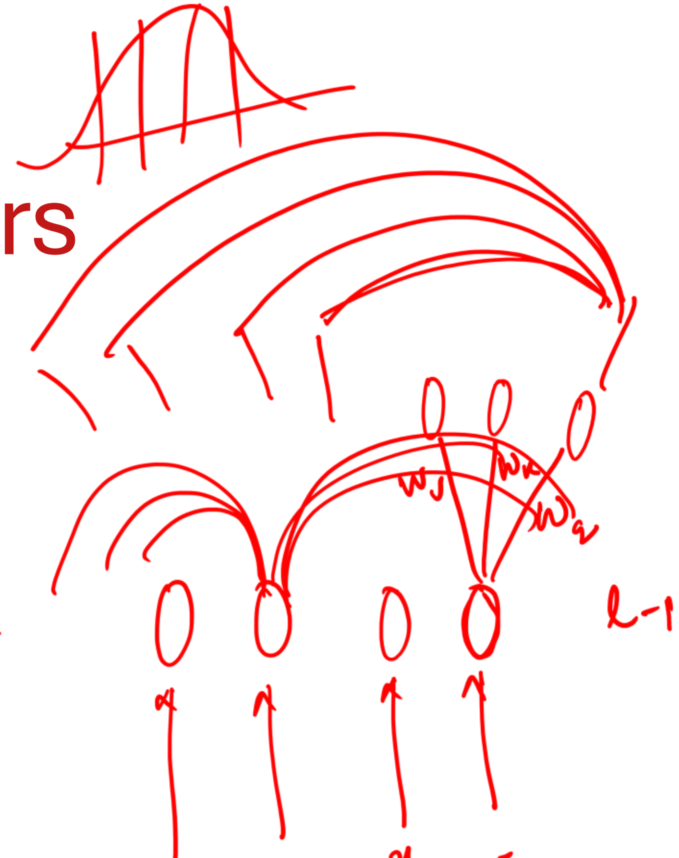
We can now get answers with links to relevant web sources!

The search model is a fine-tuned version of GPT-4o, **post-trained using novel synthetic data generation techniques, including distilling outputs from OpenAI o1-preview.** ChatGPT search leverages third-party search providers to provide the information users are looking for.



OpenAI has partnered with news and data providers to add up-to-date information for categories like weather, stocks, sports, news, and maps. ChatGPT will choose to **search the web based on what we ask**, or **we can manually choose to search by clicking the web search icon.**

Recall: Masked Self-Attention in Decoders



Self-Attention: Scaled dot-product attention

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

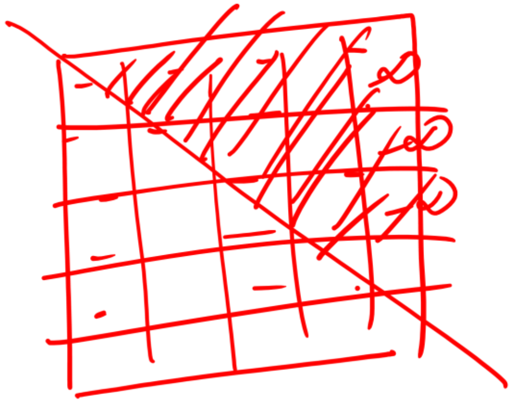
$$\text{where, } Q = XW^Q, K = XW^K, V = XW^V$$

Problem: While training autoregressive models (with next-word-prediction objective), Transformers can see the future.

- For a current token x_i, the attention scores are computed with all tokens in the sequence including those which comes after x_i (as the whole sequence is available to us during training).

Solution: Masking

Recall: Masked Self-Attention in Decoders



Masking: 'Masked' scaled dot-product attention

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

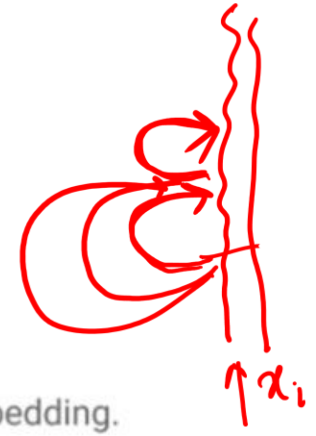
where, masking matrix M is defined as:

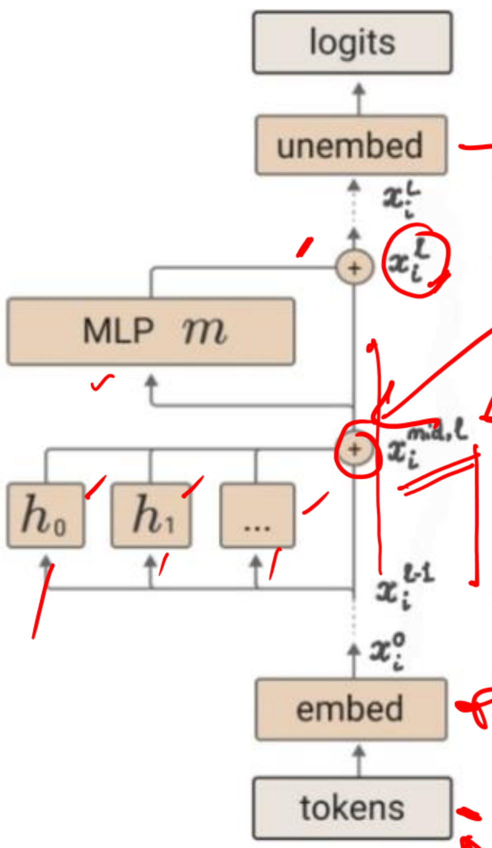
$$M_{ij} = \begin{cases} 0 & \text{if } j \leq i \\ -\infty & \text{if } j > i \end{cases}$$

For future tokens, the attention scores become zero after applying softmax [*softmax(-∞) = 0*].

- Effectively, **after masking**, the query is the current token x_i, and the keys and values come from the tokens before it, including itself (i.e., $x_j,\ j \leq i$).

Residual Stream Perspective





The final logits are produced by applying the unembedding.

$$T(t) = W_U x_i^L$$

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

$$x_i^l = x_i^{mid,l} + m(x_i^{mid,l})$$

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

$$x_i^{mid,l} = x_i^{l-1} + \sum_{h=1}^{H} \mathrm{Attn}^{l,h}(X_{\leq i}^{l-1})$$

Token embedding.

$$x_i^0 = W_E t_i$$

One residual block

Elhage, et al., A Mathematical Framework for Transformer Circuits

- Each input embedding gets updated via vector additions from the attention and feed-forward blocks producing **residual stream states** (or intermediate representations).
- The final layer residual stream state is then projected into the vocabulary space via the unembedding matrix $W_U \in R^{d \times |V|}$ and normalized via the *softmax*.

$X = \{x_1, x_2 \ldots x_n\}$

$x_1 \; x_2 \; \ldots \; x_n$ $\uparrow W_E$

$t_1 \; t_2 \ldots t_m$

$X_{n \times d}$

$x_i \in \mathbb{R}^d$

$x^l \leftarrow x_i^l \leftarrow$ ith ~~x~~ token at lth layer

$C \; X^l_{n \times d}$

$X^l_{\leq i}$ $i \times d$

$W_{OV}^{l,h} = W_V^{l,h} W_O^{l,h}$ Output-Value Circuit

Attention Block

$a_i^{l,h} = \text{Softmax}\left(x_i^{l-1} W_Q^{l,h} \left(X_{\leq i}^{l-1} W_K^{l,h} \right)^T \right)$ ⋮ $(h)\, d_h$ $d_h \neq d$

$1 \times d$ $\quad d \times d_h$ $\quad i \times d$ $\quad d \times d_h$

$1 \times d_h$ $\quad (i \times d_h)^T$

$d_h \times i$

score after QK dot product

ith X_i^{l-1}

$1 \times i$

$= \left(x_i^{l-1} \underbrace{W_Q^{l,h} W_K^{l,h^T}} X_{\leq i}^{l-1^T} \right)$

$a_i^{l,h} = [\quad]$

$W_{QK}^{l,h} = W_Q^{l,h} (W_K^{l,h})^T$ = Query-Key Circuit

$\text{Attn}^{l,h}(X_{\leq i}^{l-1}) = \sum_{j \leq i} a_{i,j}^{l,h} \left(x_j^{l-1} W_V^{l,h} \right) W_O^{l,h}$

$1 \times d \quad d \times d_h \quad d_h \times d$

$a_{i,j}^{l,h}$

$1 \times d_h$

$1 \times d$

$\text{Attn}^l(X_{\leq i}^{l-1}) = \sum_{h=1}^{H} \text{Attn}^{l,h}(X_{\leq i}^{l-1})$

$x_i^{l-1} \xrightarrow{+} x_i^{mid,l}$ d

Re-writing the Masked Self-Attention Equation

Now let's re-write the masked attention equation for a current token x_i.

- Assume that we are considering the **attention head *h*** of **layer *l*.**
- Let's denote the matrix with the **output hidden representation from layer *k* of previous tokens** $\boldsymbol{x_j},\ \boldsymbol{j \leq i}$ as $\boldsymbol{X^k_{\leq i}}$.

Thus, for calculating attention scores for **attention head *h*** of **layer *l***, input to the attention sub-layer is the output representation from the previous layer *l-1*.

- **Query:** $x_i^{l-1}\ W_Q^{l,h}$
- **Keys:** $X_{\leq i}^{l-1}\ W_K^{l,h}$

$$\boldsymbol{a}_i^{l,h} = \text{softmax}\left(\frac{\overbrace{\boldsymbol{x}_i^{l-1}\boldsymbol{W}_Q^{l,h}}^{\text{Query vector}} \underbrace{(\boldsymbol{X}_{\leq i}^{l-1}\boldsymbol{W}_K^{l,h})^\intercal}_{\text{Key vector}}}{\sqrt{d_k}} \right)$$

Ferrando et al., A Primer on the Inner Workings of Transformer-based Language Models

QK Circuit

$$\boldsymbol{a}_i^{l,h} = \text{softmax}\left(\frac{\overbrace{\boldsymbol{x}_i^{l-1}\boldsymbol{W}_Q^{l,h}}^{\text{Query vector}}\underbrace{(\boldsymbol{X}_{\leq i}^{l-1}\boldsymbol{W}_K^{l,h})^\intercal}_{\text{Key vector}}}{\sqrt{d_k}}\right)$$

$$= \text{softmax}\left(\frac{\boldsymbol{x}_i^{l-1}\boldsymbol{W}_{QK}^{h}\boldsymbol{X}_{\leq i}^{l-1\intercal}}{\sqrt{d_k}}\right),$$

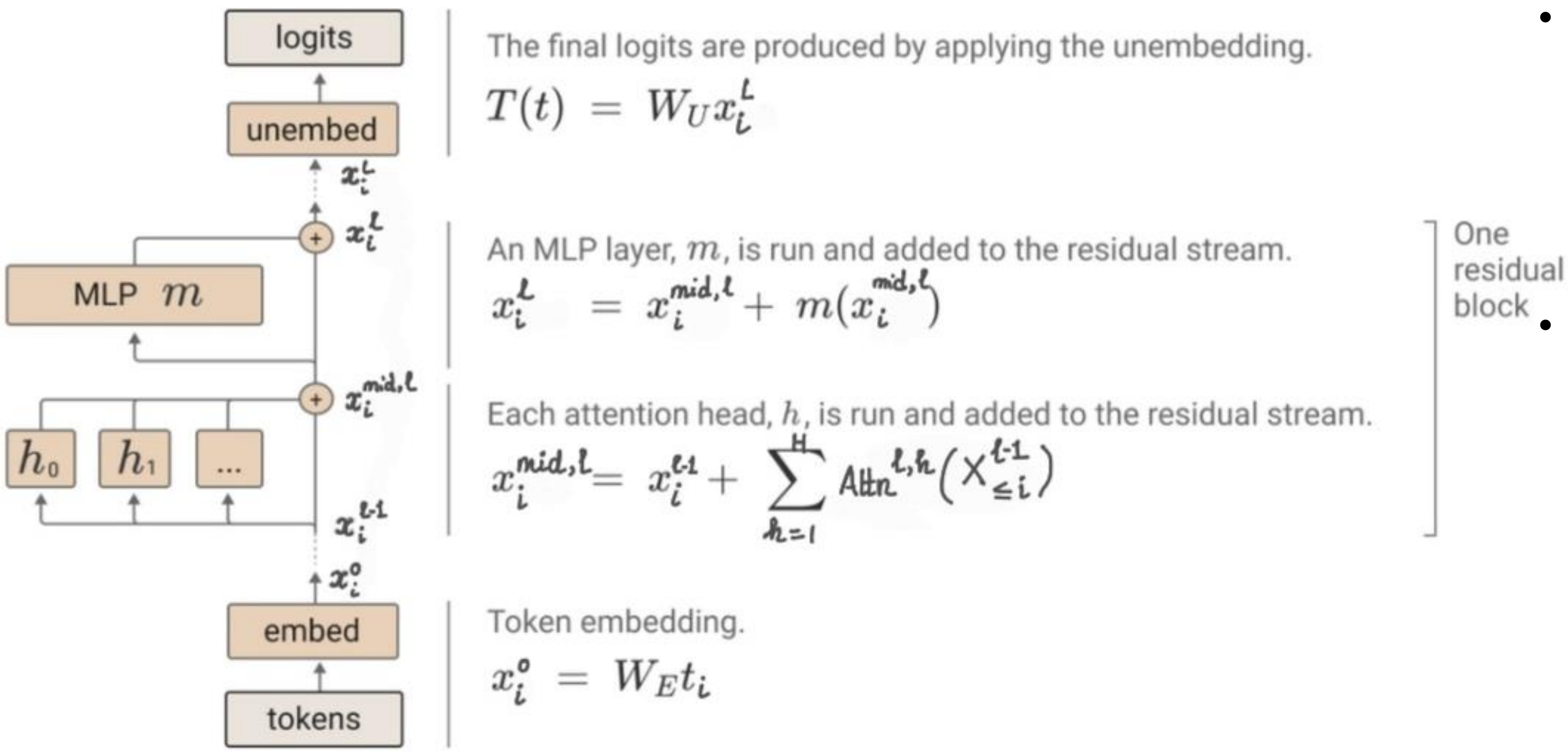
QK (query-key) circuit: $W_{QK}^h = W_Q^h W_K^{h\intercal}$

- QK circuits are responsible for reading from the **residual stream**.

Let's now look at the residual stream

Ferrando et al., A Primer on the Inner Workings of Transformer-based Language Models

Residual Stream Perspective



Elhage, et al., A Mathematical Framework for Transformer Circuits

- Each input embedding gets updated via vector additions from the attention and feed-forward blocks producing **residual stream states** (or intermediate representations).
- The final layer residual stream state is then projected into the vocabulary space via the unembedding matrix $W_U \epsilon\ R^{d \times |V|}$ and normalized via the *softmax*.

Combining the Output of Multiple Attention Heads

$$\boldsymbol{a}_i^{l,h} = \text{softmax}\left(\frac{\overbrace{\boldsymbol{x}_i^{l-1}\boldsymbol{W}_Q^{l,h}}^{\text{Query vector}}\underbrace{(\boldsymbol{X}_{\leq i}^{l-1}\boldsymbol{W}_K^{l,h})^\intercal}_{\text{Key vector}}}{\sqrt{d_k}}\right)$$

$$= \text{softmax}\left(\frac{\boldsymbol{x}_i^{l-1}\boldsymbol{W}_{QK}^{h}\boldsymbol{X}_{\leq i}^{l-1\intercal}}{\sqrt{d_k}}\right),$$

$$\text{Attn}^{l,h}(\boldsymbol{X}_{\leq i}^{l-1}) = \sum_{j \leq i} a_{i,j}^{l,h}\,\overbrace{\boldsymbol{x}_j^{l-1}\boldsymbol{W}_V^{l,h}}^{\text{Value vector}}\,\boldsymbol{W}_O^{l,h}$$

$$= \sum_{j \leq i} a_{i,j}^{l,h}\boldsymbol{x}_j^{l-1}\boldsymbol{W}_{OV}^{l,h},$$

Ferrando et al., A Primer on the Inner Workings of Transformer-based Language Models

OV Circuit

Value vector

$$\mathrm{Attn}^{l,h}(\boldsymbol{X}_{\leq i}^{l-1}) = \sum_{j \leq i} a_{i,j}^{l,h} \underbrace{\boldsymbol{x}_j^{l-1} \boldsymbol{W}_V^{l,h}}_{} \boldsymbol{W}_O^{l,h}$$

$$= \sum_{j \leq i} a_{i,j}^{l,h} \boldsymbol{x}_j^{l-1} \boldsymbol{W}_{OV}^{l,h},$$

OV (output-value) circuit: $W_{OV}^{l,h} = W_V^{l,h} W_O^{l,h}$

- OV circuits are responsible for writing to the **residual stream**.

Ferrando et al., A Primer on the Inner Workings of Transformer-based Language Models

Attention Block Output

The attention block output is the sum of individual attention heads, which is subsequently added back into the residual stream.

$$\mathrm{Attn}^{l,h}(\boldsymbol{X}_{\leq i}^{l-1}) = \sum_{j \leq i} a_{i,j}^{l,h} \underbrace{\boldsymbol{x}_j^{l-1} \boldsymbol{W}_V^{l,h}}_{\text{Value vector}} \boldsymbol{W}_O^{l,h} = \sum_{j \leq i} a_{i,j}^{l,h} \boldsymbol{x}_j^{l-1} \boldsymbol{W}_{OV}^{l,h},$$

$$\boldsymbol{a}_i^{l,h} = \mathrm{softmax}\left(\frac{\underbrace{\boldsymbol{x}_i^{l-1} \boldsymbol{W}_Q^{l,h}}_{\text{Query vector}} \underbrace{(\boldsymbol{X}_{\leq i}^{l-1} \boldsymbol{W}_K^{l,h})^{\intercal}}_{\text{Key vector}}}{\sqrt{d_k}} \right) = \mathrm{softmax}\left(\frac{\boldsymbol{x}_i^{l-1} \boldsymbol{W}_{QK}^{h} \boldsymbol{X}_{\leq i}^{l-1\intercal}}{\sqrt{d_k}} \right),$$

$$\boxed{\begin{aligned} \mathrm{Attn}^{l}(\boldsymbol{X}_{\leq i}^{l-1}) &= \sum_{h=1}^{H} \mathrm{Attn}^{l,h}(\boldsymbol{X}_{\leq i}^{l-1}) \\ \boldsymbol{x}_i^{\mathrm{mid},l} &= \boldsymbol{x}_i^{l-1} + \mathrm{Attn}^{l}(\boldsymbol{X}_{\leq i}^{l-1}). \end{aligned}}$$

Ferrando et al., A Primer on the Inner Workings of Transformer-based Language Models

Feed-Forward Network (FFN)

$1 \times d_f$

d_f

$1 \times d$

$d \times d_f$

$d_f \times d$

$$\mathrm{FFN}^l(\boldsymbol{x}_i^{\mathrm{mid},l}) = g(\boldsymbol{x}_i^{\mathrm{mid},l}\boldsymbol{W}_{\mathrm{in}}^l)\boldsymbol{W}_{\mathrm{out}}^l.$$

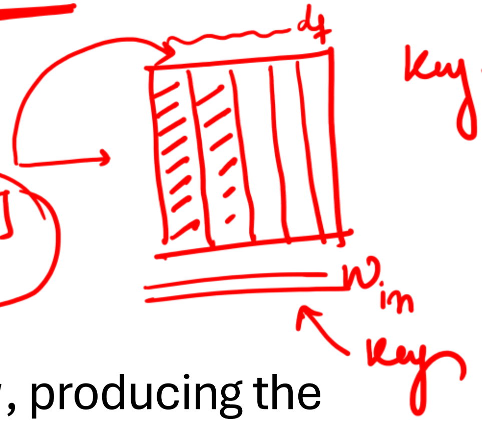
$$\boldsymbol{x}_i^l = \boldsymbol{x}_i^{\mathrm{mid},l} + \mathrm{FFN}^l(\boldsymbol{x}_i^{\mathrm{mid},l}).$$

$$\boldsymbol{W}_{\mathrm{in}}^l \in \mathbb{R}^{d \times d_{\mathrm{ffn}}}$$

$$\boldsymbol{W}_{\mathrm{out}}^l \in \mathbb{R}^{d_{\mathrm{ffn}} \times d}$$

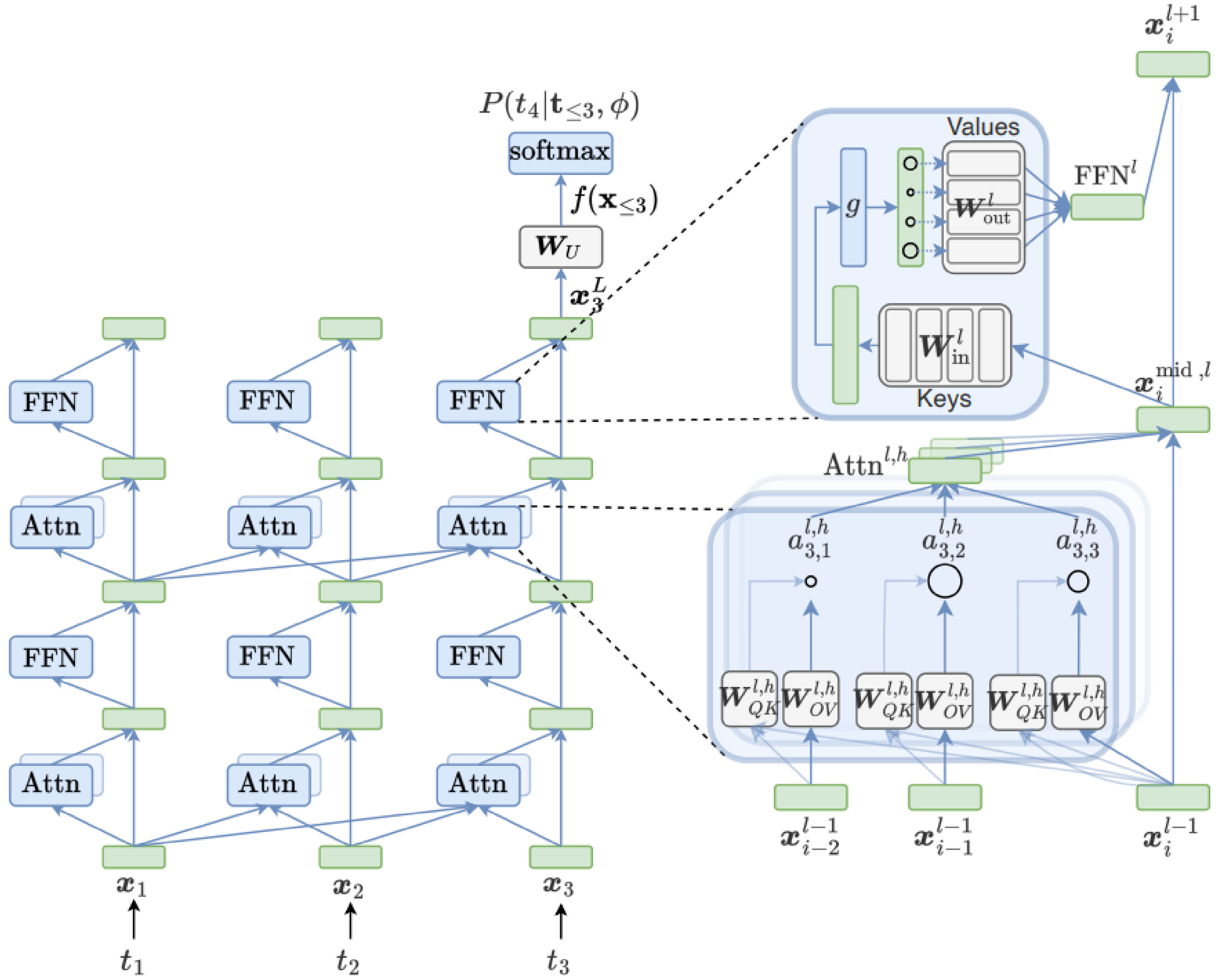
key-value memory retrieval

$[0\,0\,1]$ x_i

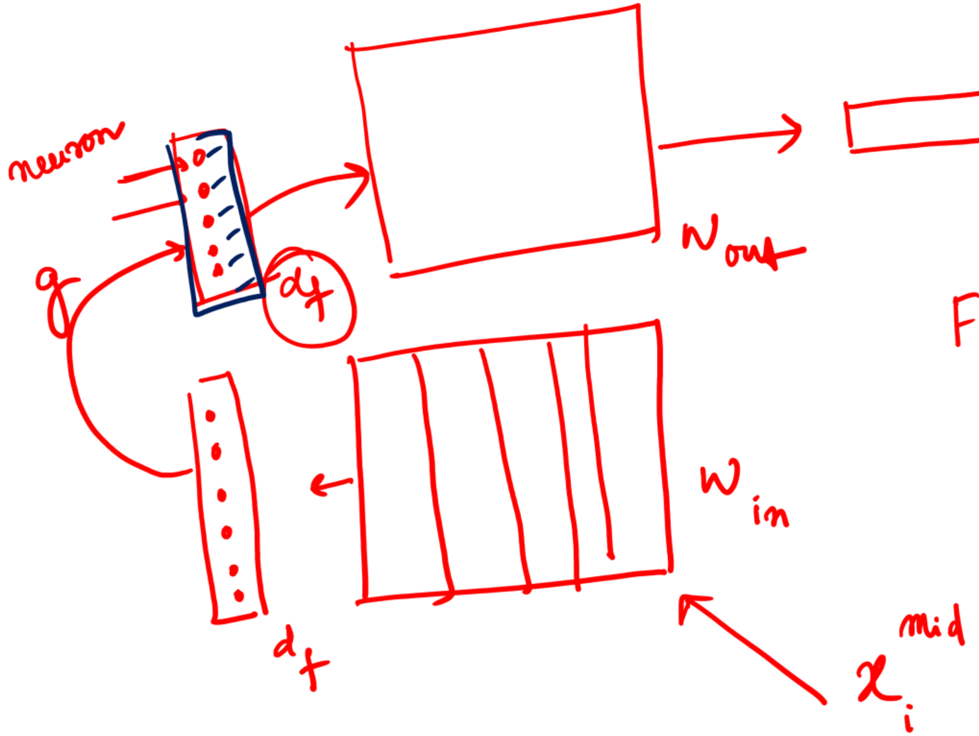


- W_{in}^l reads from the residual stream state $x_i^{mid,l}$.
- Its result is passed through an element-wise non-linear activation function g, producing the neuron activations.
- These get transformed by W_{out}^l to produce the output $FFN^l(x_i^{mid,l})$, which is then added back to the residual stream

Ferrando et al., A Primer on the Inner Workings of Transformer-based Language Models

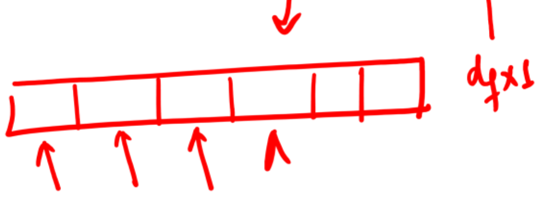


$P(t_4|\mathbf{t}_{\leq 3}, \phi)$
softmax
$f(\mathbf{x}_{\leq 3})$
W_U
x_3^L
FFN
Attn
x_1
x_2
x_3
t_1
t_2
t_3
Values
g
W_{out}^l
FFN^l
x_i^{l+1}
W_{in}^l
Keys
$x_i^{mid,l}$
$Attn^{l,h}$
$a_{3,1}^{l,h}$
$a_{3,2}^{l,h}$
$a_{3,3}^{l,h}$
$W_{QK}^{l,h}$
$W_{OV}^{l,h}$
x_{i-2}^{l-1}
x_{i-1}^{l-1}
x_i^{l-1}



$$FFN^{l}\left(x_i^{mid,l}\right) = \sum_{u=1}^{d_f} \underbrace{g_u\left(\underline{x_i^{mid,l}}\, W_{in_u}^{l}\right)}_{n_u^l} W_{out_u}^{l}$$

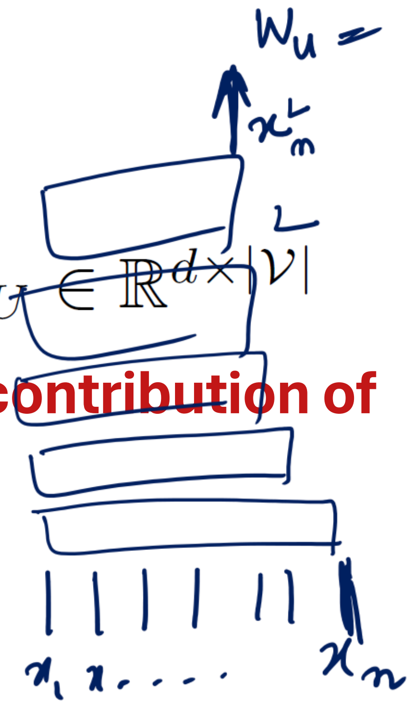
$1 \times d$ $\quad$ $d \times 1$



n_u^l : $n^l \in R^{d_f}$

vector of neuron activation

Prediction as a Sum of Component Outputs



- Prediction head of a Transformer consists of an unembedding matrix: $\boldsymbol{W}_U \in \mathbb{R}^{d \times |\mathcal{V}|}$

We can rearrange the traditional forward pass formulation to separate the **contribution of each model component to the output logits**: L

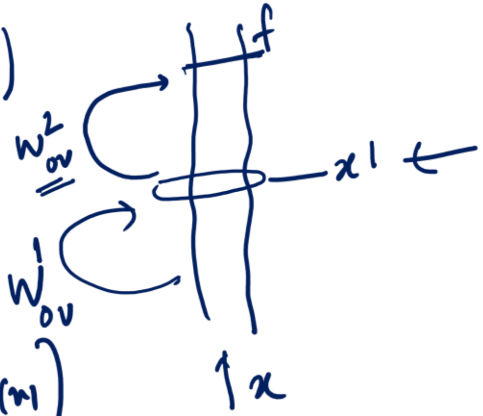
$$f(\mathbf{x}) = \boldsymbol{x}_n^L \boldsymbol{W}_U$$

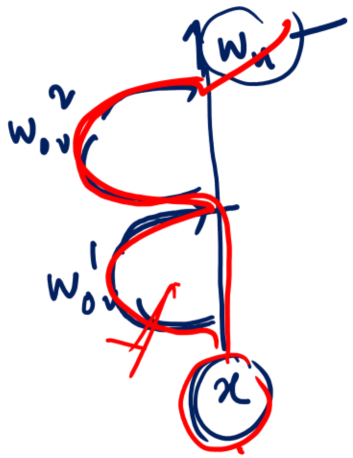
$$= \left(\sum_{l=1}^{L} \sum_{h=1}^{H} \mathrm{Attn}^{l,h}(\boldsymbol{X}_{\leq n}^{l-1}) + \sum_{l=1}^{L} \mathrm{FFN}^{l}(\boldsymbol{x}_n^{\mathrm{mid},l}) + \boldsymbol{x}_n \right) \boldsymbol{W}_U$$

$$= \sum_{l=1}^{L} \sum_{h=1}^{H} \underbrace{\mathrm{Attn}^{l,h}(\boldsymbol{X}_{\leq n}^{l-1}) \boldsymbol{W}_U}_{\text{Attention head logits update}} + \sum_{l=1}^{L} \underbrace{\mathrm{FFN}^{l}(\boldsymbol{x}_n^{\mathrm{mid},l}) \boldsymbol{W}_U}_{\text{FFN logits update}} + \boldsymbol{x}_n \boldsymbol{W}_U .$$

Ferrando et al., A Primer on the Inner Workings of Transformer-based Language Models

$$x' = x + W_{OV}^1(x)$$



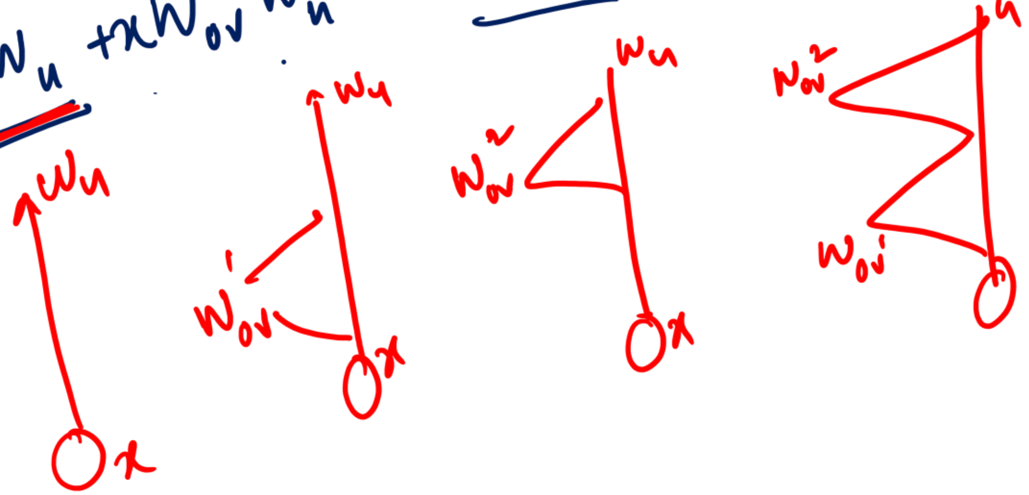


$$f(x) = x' + W_{OV}^2(x')$$

$$= x + W_{OV}^1(x) + W_{OV}^2\left(x + W_{OV}^1(x)\right)$$

$$= \left(x + W_{OV}^1 x + W_{OV}^2 x + W_{OV}^1 W_{OV}^2 x\right) W_U$$

$$= x W_U + x W_{OV}^1 W_U + x W_{OV}^2 W_U + x W_{OV}^1 W_{OV}^2 W_U$$



Prediction as an Ensemble of Shallow Networks

- Residual networks work as ensembles of shallow networks, where **each subnetwork defines a path in the computational graph**.

Consider a two-layer attention-only Transformer, where each attention head is composed just by an OV matrix:

$$f(\boldsymbol{x}) = \boldsymbol{x}^1 + \boldsymbol{W}_{OV}^2(\boldsymbol{x}^1), \text{ with } \boldsymbol{x}^1 = \boldsymbol{x} + \boldsymbol{W}_{OV}^1(\boldsymbol{x})$$

We can decompose the forward pass as:

Direct path

Full OV circuits

$$f(\boldsymbol{x}) = \boldsymbol{x}\boldsymbol{W}_U + \boldsymbol{x}\boldsymbol{W}_{OV}^1\boldsymbol{W}_U + \boldsymbol{x}\boldsymbol{W}_{OV}^1\boldsymbol{W}_{OV}^2\boldsymbol{W}_U + \boldsymbol{x}\boldsymbol{W}_{OV}^2\boldsymbol{W}_U .$$

Virtual attention heads (V-composition)

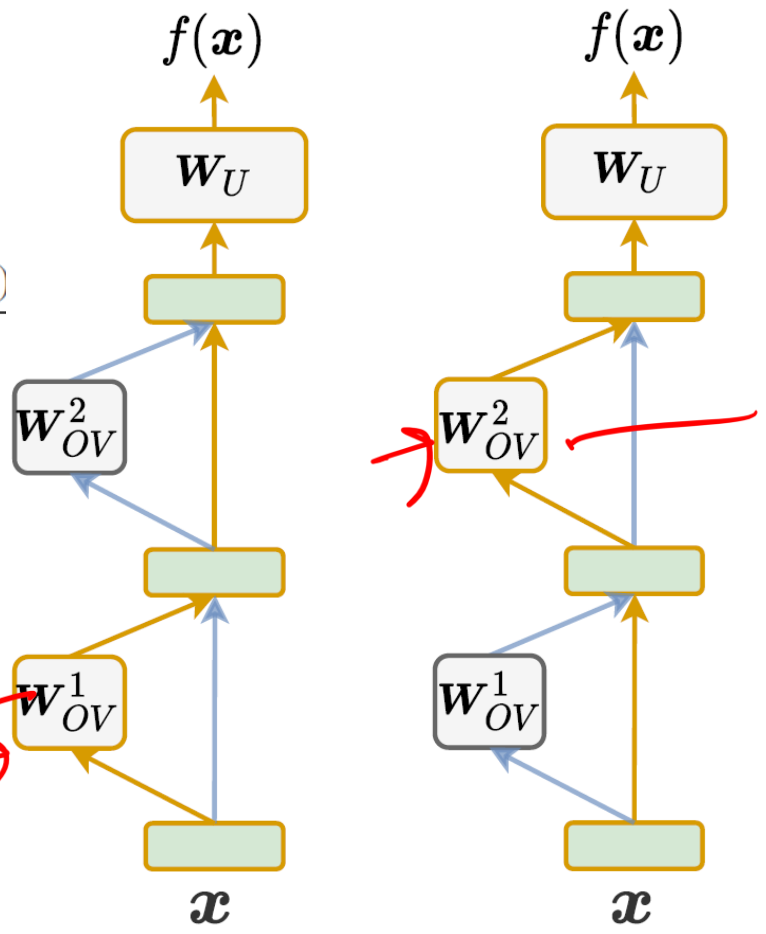
Ferrando et al., A Primer on the Inner Workings of Transformer-based Language Models

Prediction as an Ensemble of Shallow Networks

$$f(\boldsymbol{x}) = \overbrace{\boldsymbol{x}\boldsymbol{W}_U}^{\text{Direct path}} + \boldsymbol{x}\boldsymbol{W}_{OV}^1\boldsymbol{W}_U + \underbrace{\boldsymbol{x}\boldsymbol{W}_{OV}^1\boldsymbol{W}_{OV}^2\boldsymbol{W}_U}_{\text{Virtual attention heads (V-composition)}} + \boldsymbol{x}\boldsymbol{W}_{OV}^2\boldsymbol{W}_U.$$

Full OV circuits: $\boldsymbol{x}\boldsymbol{W}_{OV}^1\boldsymbol{W}_U$, $\boldsymbol{x}\boldsymbol{W}_{OV}^2\boldsymbol{W}_U$

- These terms depict paths traversing a single OV circuit, and are named **full OV circuits**
- The contribution of each OV circuit towards the output logit of the next token to be predicted.



Ferrando et al., A Primer on the Inner Workings of Transformer-based Language Models

Prediction as an Ensemble of Shallow Networks

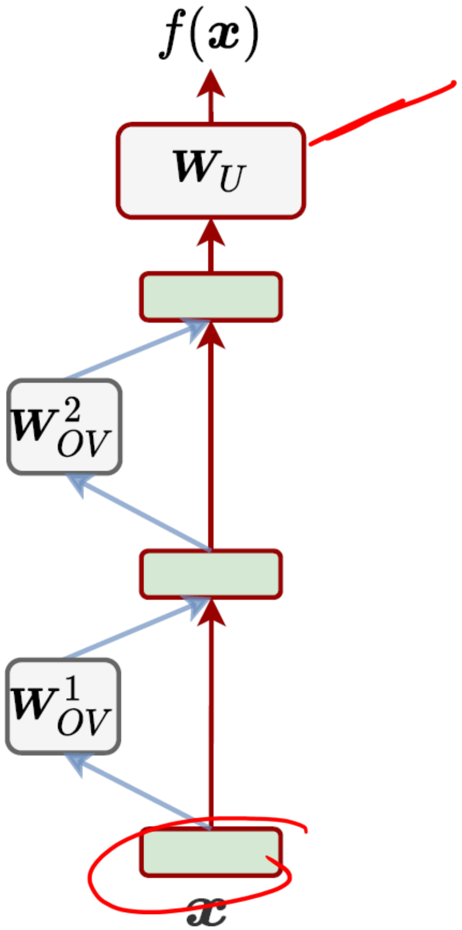
Direct path

Full OV circuits

$$f(\boldsymbol{x}) = \boldsymbol{x}\boldsymbol{W}_U + \boldsymbol{x}\boldsymbol{W}_{OV}^{1}\boldsymbol{W}_U + \boldsymbol{x}\boldsymbol{W}_{OV}^{1}\boldsymbol{W}_{OV}^{2}\boldsymbol{W}_U + \boldsymbol{x}\boldsymbol{W}_{OV}^{2}\boldsymbol{W}_U .$$

Virtual attention heads (V-composition

W_{OV}^{12}



- This term links the input embedding to the unembedding matrix and is referred to as the **direct path**
- It shows the contribution of the input embedding towards the output logit of the next token to be predicted.

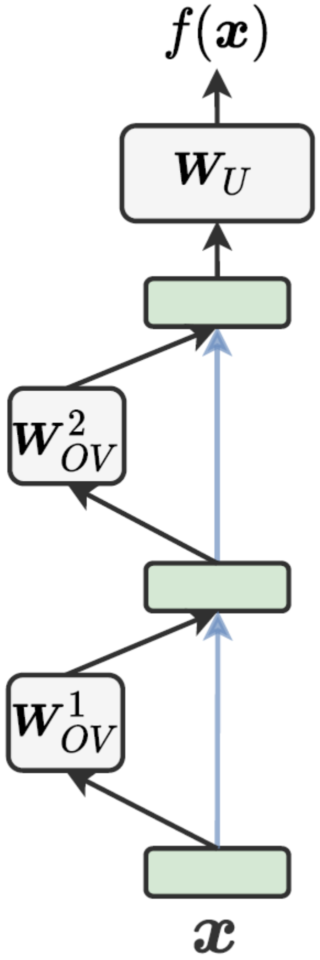
Ferrando et al., A Primer on the Inner Workings of Transformer-based Language Models

Prediction as an Ensemble of Shallow Networks

Direct path — Full OV circuits

$$f(\boldsymbol{x}) = \boldsymbol{x}\boldsymbol{W}_U + \boldsymbol{x}\boldsymbol{W}_{OV}^1\boldsymbol{W}_U + \boldsymbol{x}\boldsymbol{W}_{OV}^1\boldsymbol{W}_{OV}^2\boldsymbol{W}_U + \boldsymbol{x}\boldsymbol{W}_{OV}^2\boldsymbol{W}_U .$$

Virtual attention heads (V-composition)



- This term depicts the path involving both attention heads, and is referred to as **virtual attention heads doing V-composition**
- This is called 'composition' since the sequential writing and reading of the two heads is seen as OV matrices composing together.
  - The amount of composition can be measured as:

$$||W_{OV}^1 W_{OV}^2||_F \Big/ ||W_{OV}^1||_F ||W_{OV}^2||_F$$

Ferrando et al., A Primer on the Inner Workings of Transformer-based Language Models

Prediction as an Ensemble of Shallow Networks

- In full Transformer models, Q-composition and K-composition, i.e., compositions of W_Q and W_K with the W_{OV} output of previous layers, can also be found.

- Such decomposition enables us to localize the inputs or model components responsible for a particular prediction.

Why Do We Need Such a Formulation?

- Better understand the information flow within Transformer-based LLMs.
- Reveals how each layer incrementally transforms token representations.
  - Shows how attention heads and FFNs contribute to language modeling.
- Breaking down the contributions of individual circuits allows us to interpret which aspects of the model influence specific predictions.

Thus, through this formulation, the behavior of attention heads, the interaction between tokens, and the role of the residual stream can be explored more clearly.