Transformers & LLMs

Md. Mohsinul Kabir
Islamic University of Technology

Self Attention

Things covered so far:

Comparison:

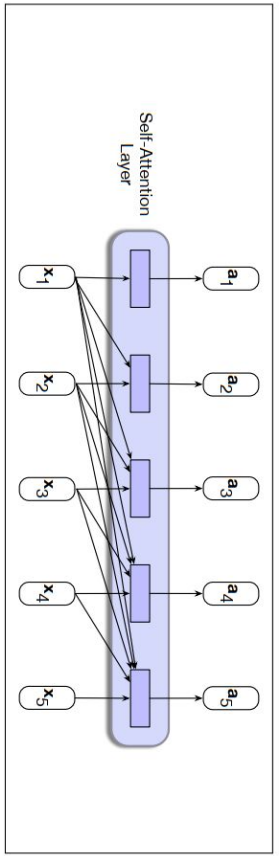
$$\text{score}(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i \cdot \mathbf{x}_j$$

Normalizing:

$$\begin{aligned} \alpha_{ij} &= \text{softmax}(\text{score}(\mathbf{x}_i, \mathbf{x}_j)) \ \forall j \leq i \\ &= \frac{\exp(\text{score}(\mathbf{x}_i, \mathbf{x}_j))}{\sum_{k=1}^{i} \exp(\text{score}(\mathbf{x}_i, \mathbf{x}_k))} \ \forall j \leq i \end{aligned}$$

Value Calculation:

$$\mathbf{a}_i = \sum_{j \leq i} \alpha_{ij} \mathbf{x}_j$$



Self Attention

Consider the three different roles that each input embedding plays during the course of the attention process:

- **Query**: As the current focus of attention when being compared to all of the other preceding inputs. We'll refer to this role as a query.

- **Key**: In its role as a preceding input being compared to the current focus of attention. We'll refer to this role as a key.

- **Value**: And finally, as a value used to compute the output for the current focus of attention.

Self Attention

Consider the three different roles that each input embedding plays during the course of the attention process:

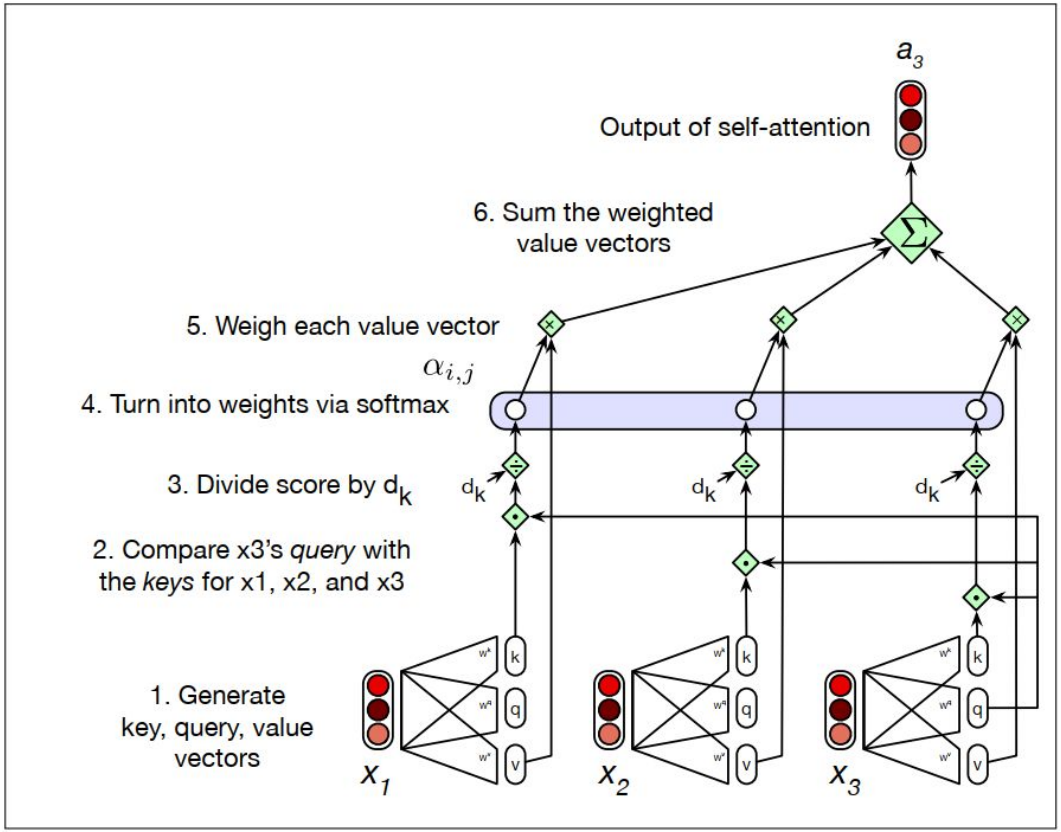
- **Query**: As the current focus of attention when being compared to all of the other preceding inputs. We'll refer to this role as a query.

- **Key**: In its role as a preceding input being compared to the current focus of attention. We'll refer to this role as a key.

- **Value**: And finally, as a value used to compute the output for the current focus of attention.

$$\mathbf{q}_i = \mathbf{x}_i\mathbf{W}^{\mathbf{Q}};\ \ \mathbf{k}_i = \mathbf{x}_i\mathbf{W}^{\mathbf{K}};\ \ \mathbf{v}_i = \mathbf{x}_i\mathbf{W}^{\mathbf{V}}$$

Self Attention



Self Attention

The result of a dot product can be an arbitrarily large (positive or negative) value. Exponentiating large values can lead to numerical issues and to an effective loss of gradients during training.

$$\mathbf{q}_i = \mathbf{x}_i\mathbf{W}^{\mathbf{Q}}; \mathbf{k}_i = \mathbf{x}_i\mathbf{W}^{\mathbf{K}}; \mathbf{v}_i = \mathbf{x}_i\mathbf{W}^{\mathbf{V}}$$

Final verson:

$$\text{score}(\mathbf{x}_i, \mathbf{x}_j) = \frac{\mathbf{q}_i \cdot \mathbf{k}_j}{\sqrt{d_k}}$$

$$\alpha_{ij} = \text{softmax}(\text{score}(\mathbf{x}_i, \mathbf{x}_j)) \ \forall j \leq i$$

$$\mathbf{a}_i = \sum_{j \leq i} \alpha_{ij}\mathbf{v}_j$$

Masking out the Future

$$\mathbf{A} = \text{SelfAttention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\intercal}{\sqrt{d_k}}\right)\mathbf{V}$$

N

q1•k1	−∞	−∞	−∞	−∞
q2•k1	q2•k2	−∞	−∞	−∞
q3•k1	q3•k2	q3•k3	−∞	−∞
q4•k1	q4•k2	q4•k3	q4•k4	−∞
q5•k1	q5•k2	q5•k3	q5•k4	q5•k5

N

Multihead Attention

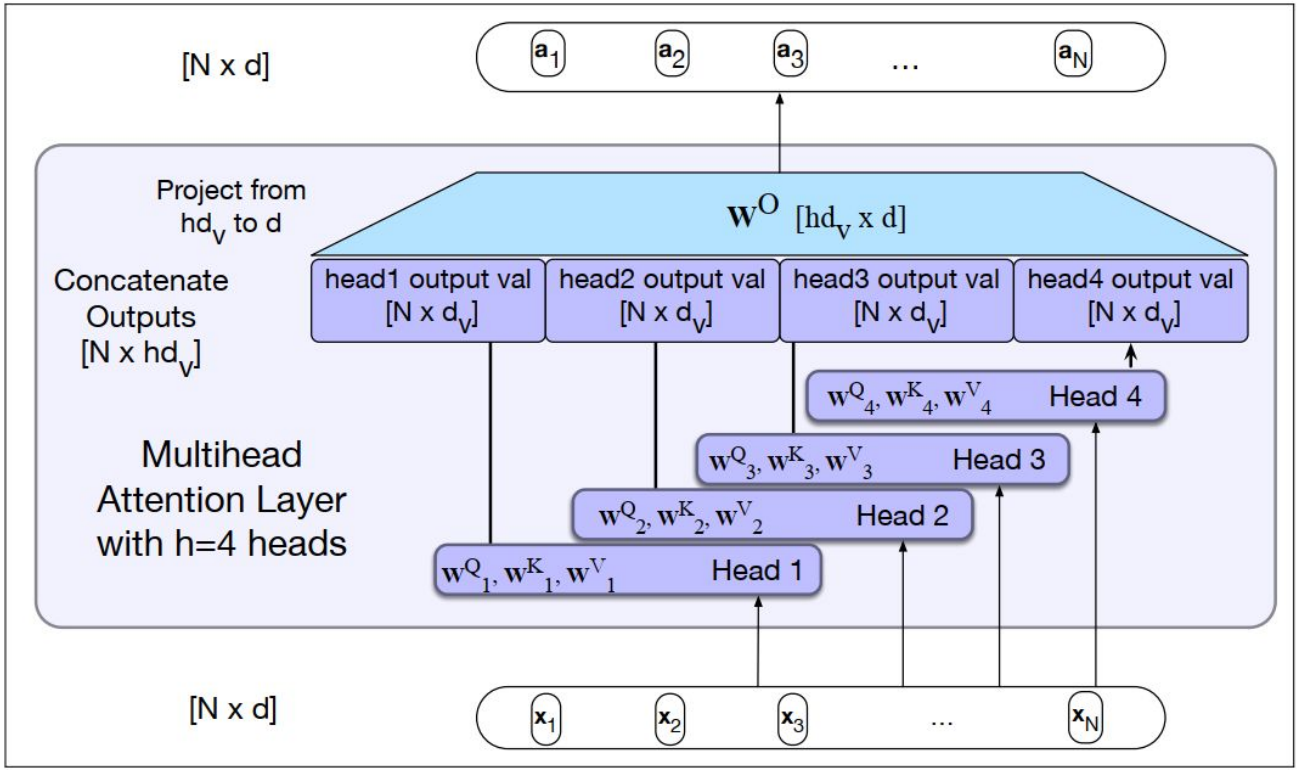
- Different words in a sentence can relate to each other in many different ways simultaneously.
- Difficult for a single self-attention model to learn to capture all of the different kinds of parallel relations among its inputs.
- **Solution? Multihead Self-Attention Layers.**

$$\mathbf{Q} = \mathbf{X}\mathbf{W}_i^Q \; ; \; \mathbf{K} = \mathbf{X}\mathbf{W}_i^K \; ; \; \mathbf{V} = \mathbf{X}\mathbf{W}_i^V$$

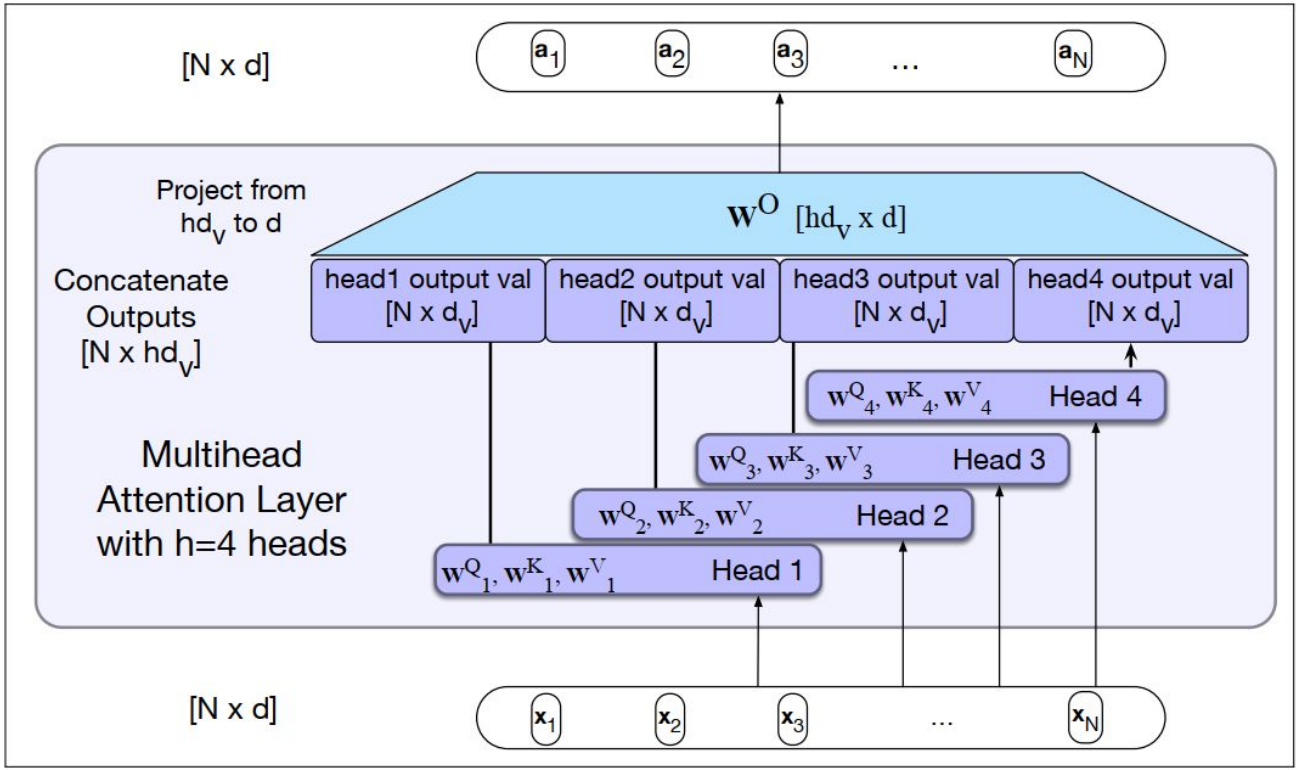
$$\mathbf{head}_i = \mathrm{SelfAttention}(\mathbf{Q}, \mathbf{K}, \mathbf{V})$$

$$\mathbf{A} = \mathrm{MultiHeadAttention}(\mathbf{X}) = (\mathbf{head}_1 \oplus \mathbf{head}_2 \ldots \oplus \mathbf{head}_h)\mathbf{W}^O$$

Multihead Attention

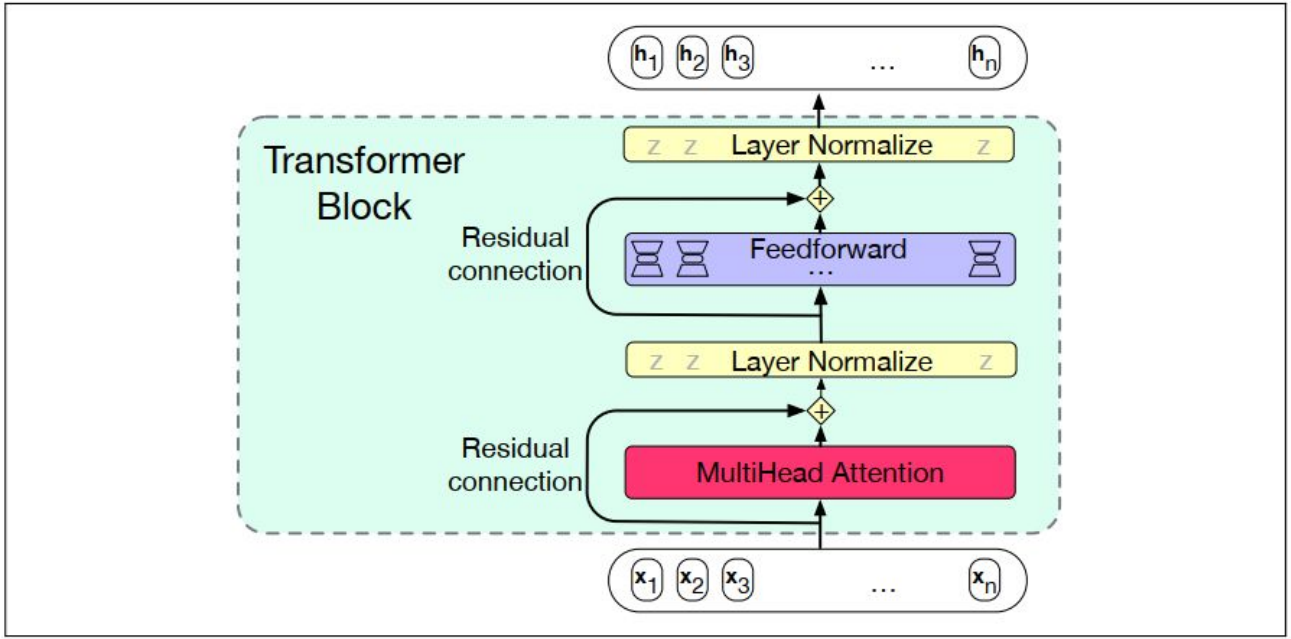


Multihead Attention



Transformer Blocks

The are **four** kinds of layers in a single transformer block: self-attention, feedforward, residual connection and normalizing layers.



Transformer Blocks

Residual Connection: passes information from a lower layer to a higher layer without going through the intermediate layer. Allowing information from the activation going forward and the gradient going backwards to skip a layer **improves learning** and gives higher level layers **direct access to information** from lower layers (He et al., 2016)

Layer Norm: improves training performance in deep neural networks by keeping the values of a hidden layer in a range that facilitates gradient-based training.

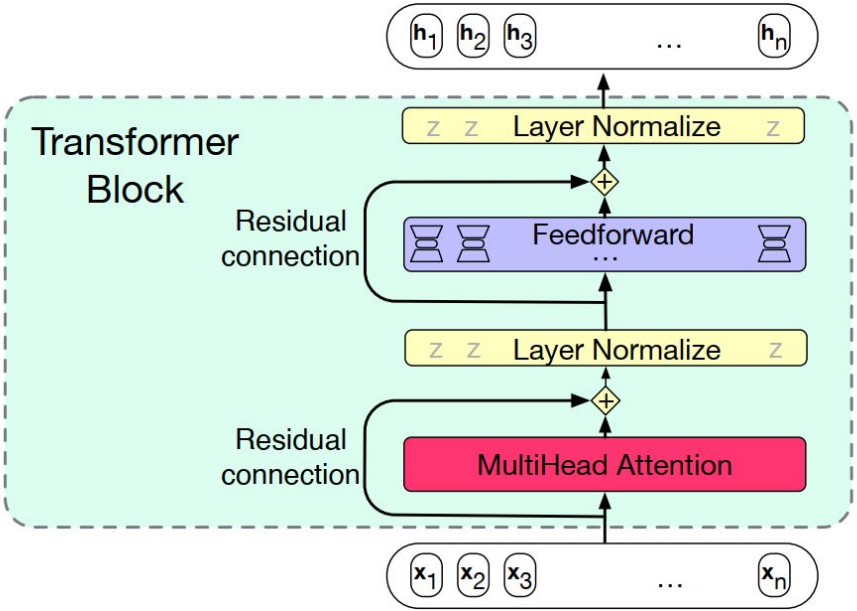
$$\mu = \frac{1}{d_h}\sum_{i=1}^{d_h} x_i$$

$$\sigma = \sqrt{\frac{1}{d_h}\sum_{i=1}^{d_h}(x_i-\mu)^2}$$

$$\hat{\mathbf{x}} = \frac{(\mathbf{x}-\mu)}{\sigma}$$

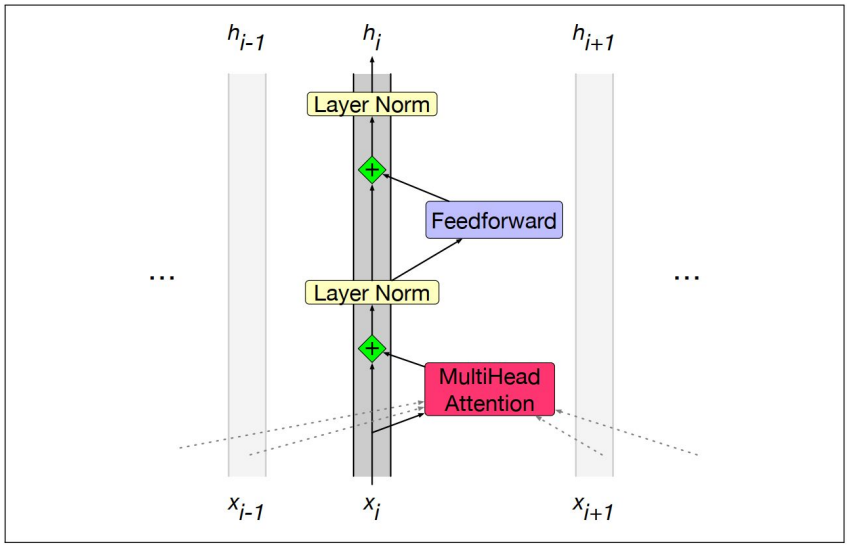
$$\text{LayerNorm} = \gamma\,\hat{\mathbf{x}} + \beta$$

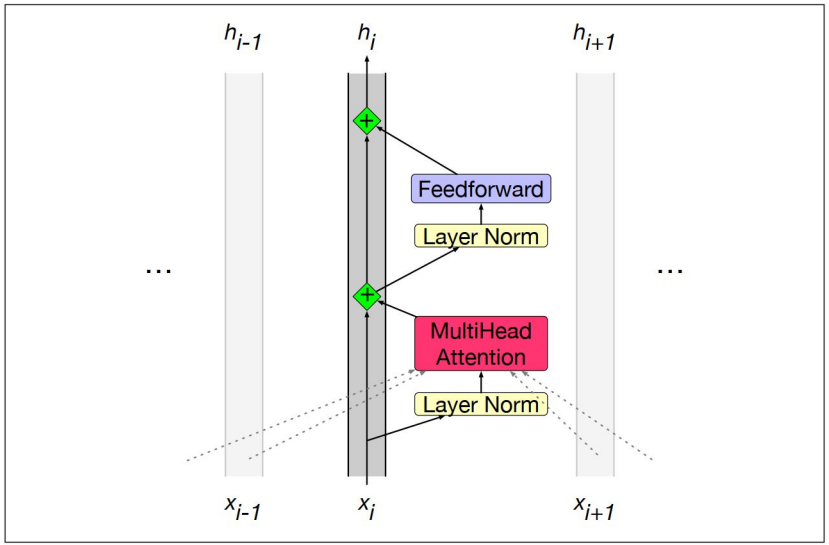
Putting it All Together



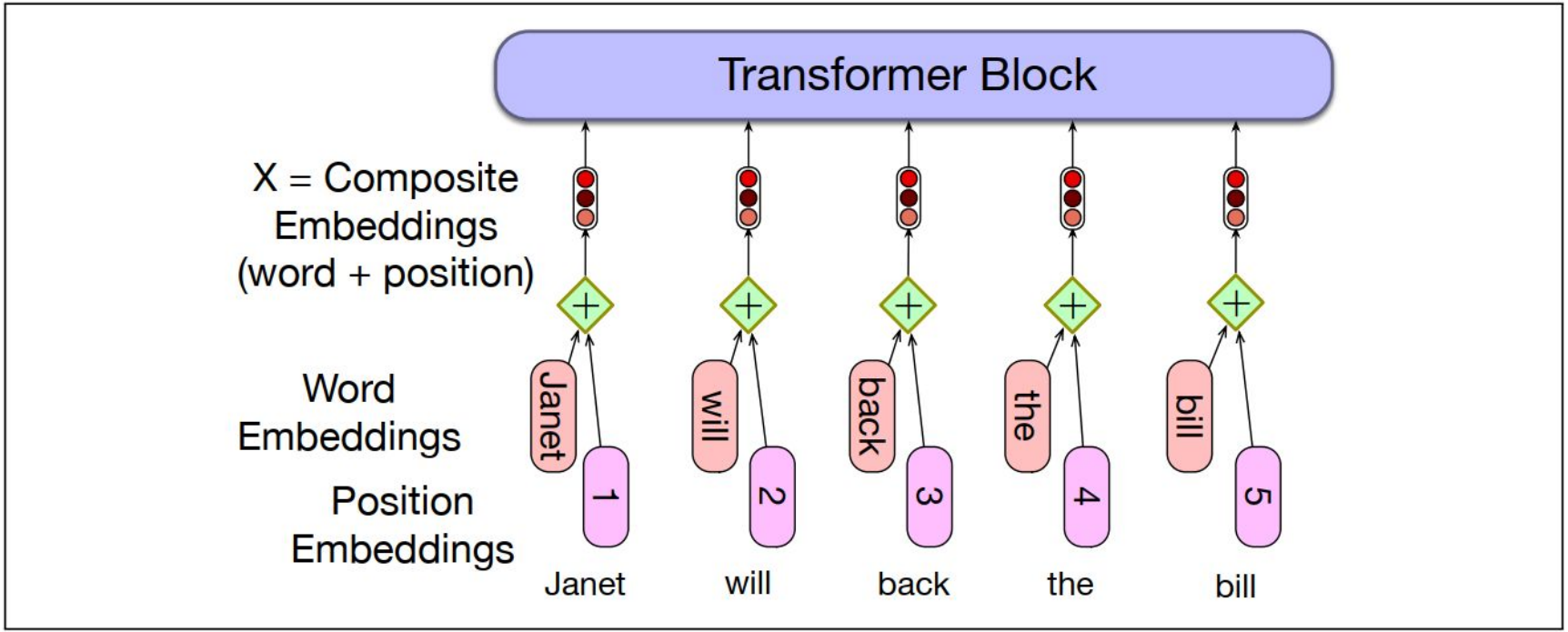
$$\mathbf{T}^1 = \text{SelfAttention}(\mathbf{X})$$
$$\mathbf{T}^2 = \mathbf{X} + \mathbf{T}^1$$
$$\mathbf{T}_3 = \text{LayerNorm}(\mathbf{T}^2)$$
$$\mathbf{T}^4 = \text{FFN}(\mathbf{T}^3)$$
$$\mathbf{T}^5 = \mathbf{T}^4 + \mathbf{T}^3$$
$$\mathbf{H} = \text{LayerNorm}(\mathbf{T}^5)$$

Pre-norm vs Post-norm Transformers

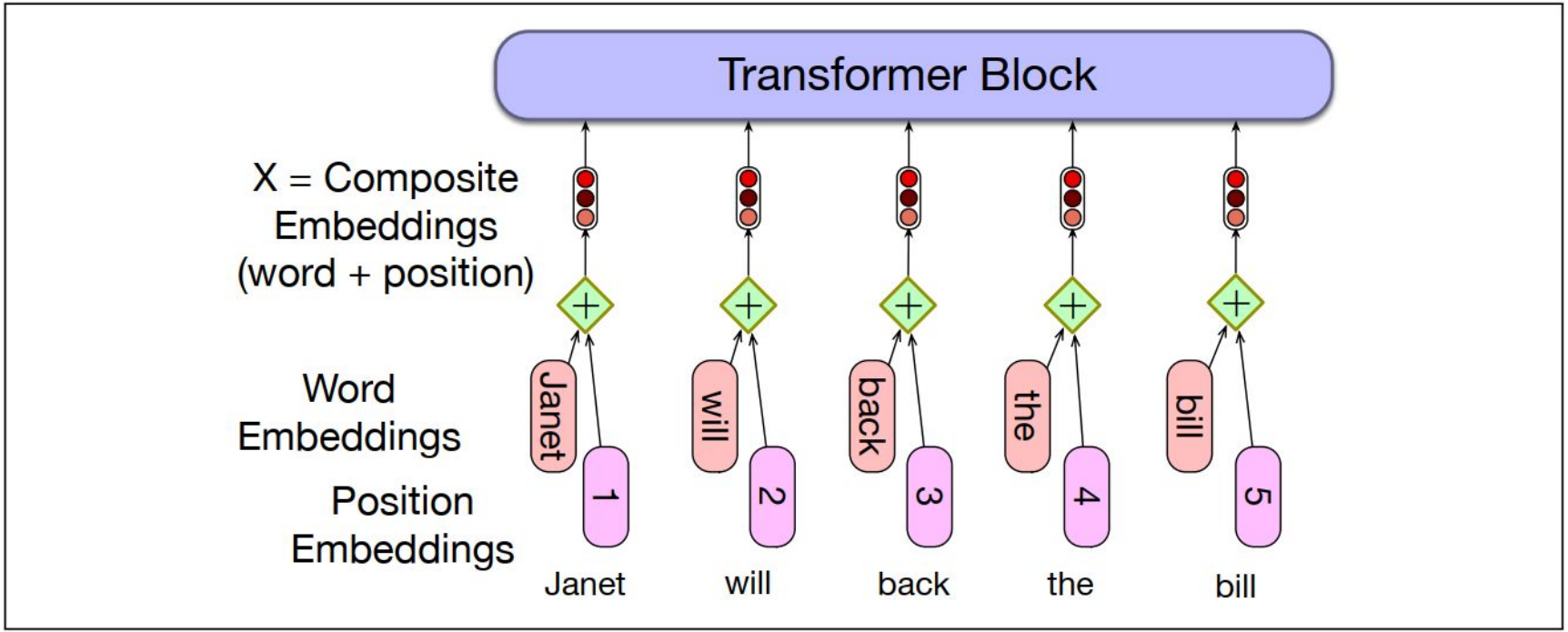




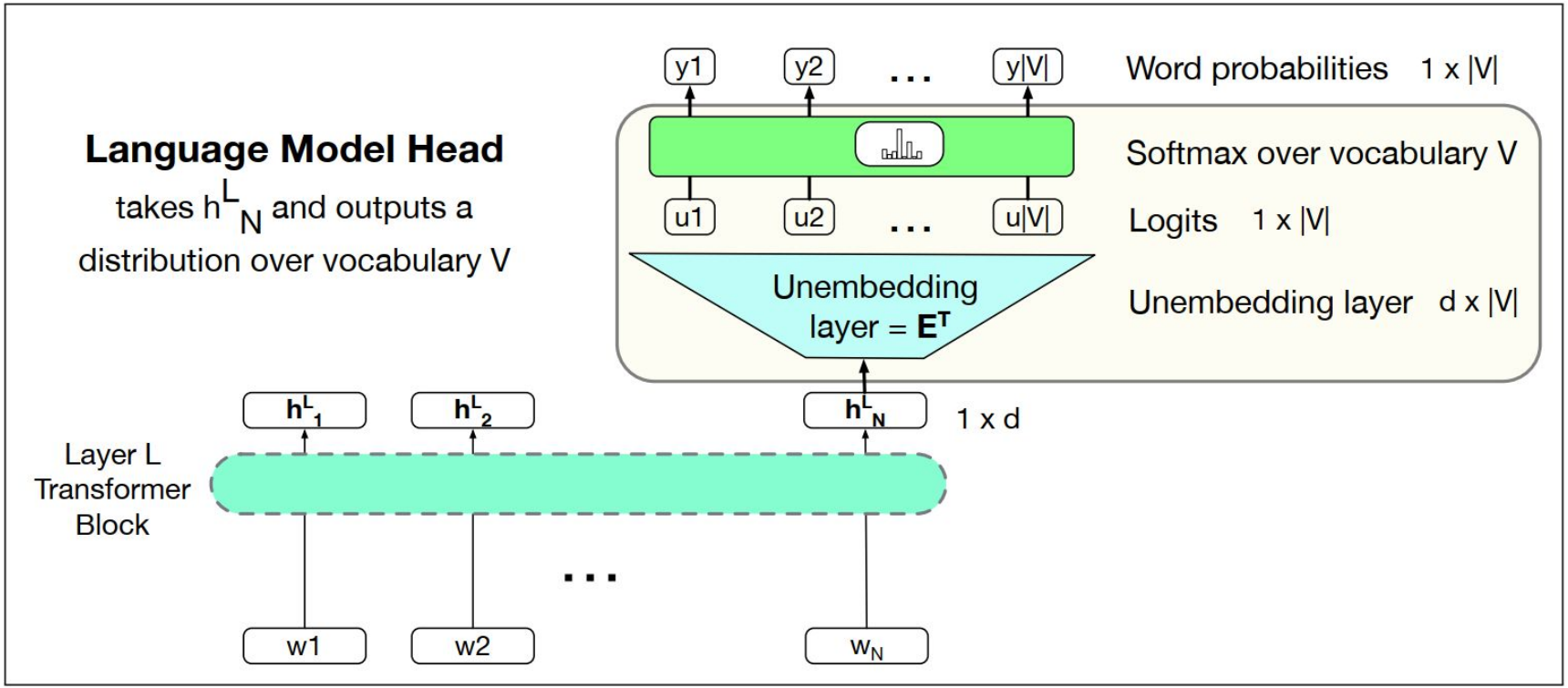
Word Embeddings & Position Embeddings



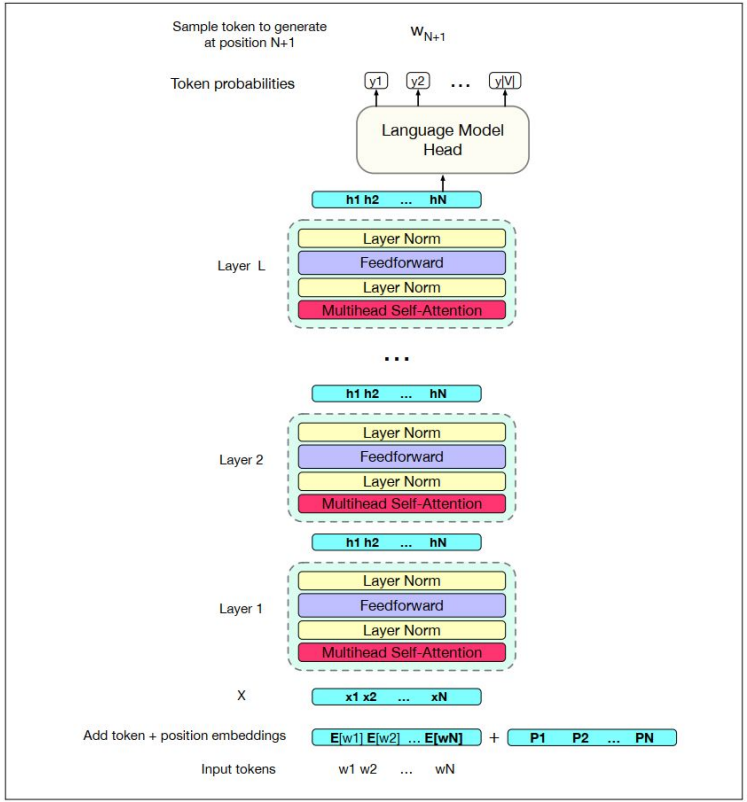
Word Embeddings & Position Embeddings



Language Modeling Head



Language Modeling Head



LLMs with Transformers

The fact that transformers have such long contexts (1024 or even 4096 tokens) makes them very powerful for conditional generation, because they can look back so far into the prompting text.

Many practical NLP tasks can be cast as word prediction.

Sentiment Analysis:

The sentiment of the sentence "I like Jackie Chan" is:

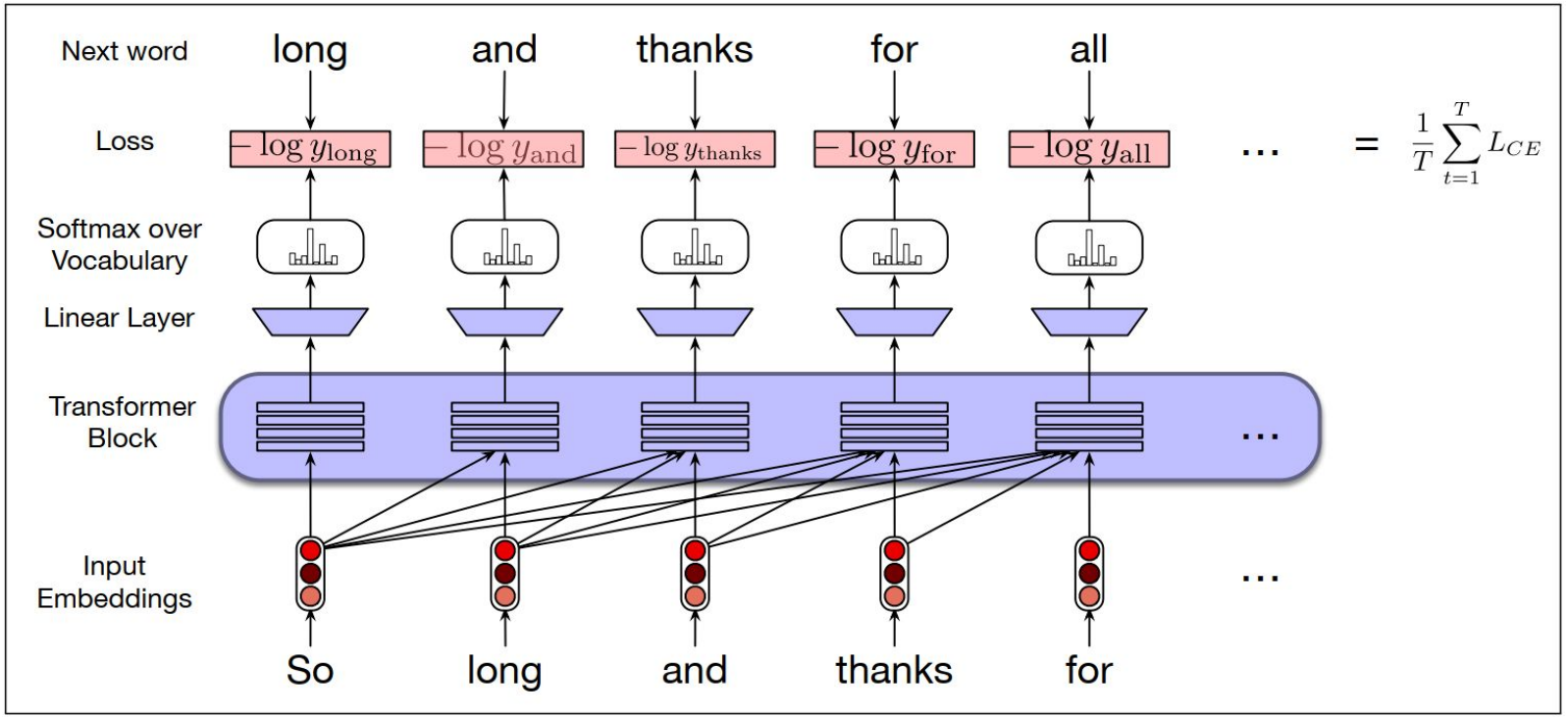
Question-Answering:

P(w |Q: Who wrote the book "The Origin of Species"? A:)

Summarization:

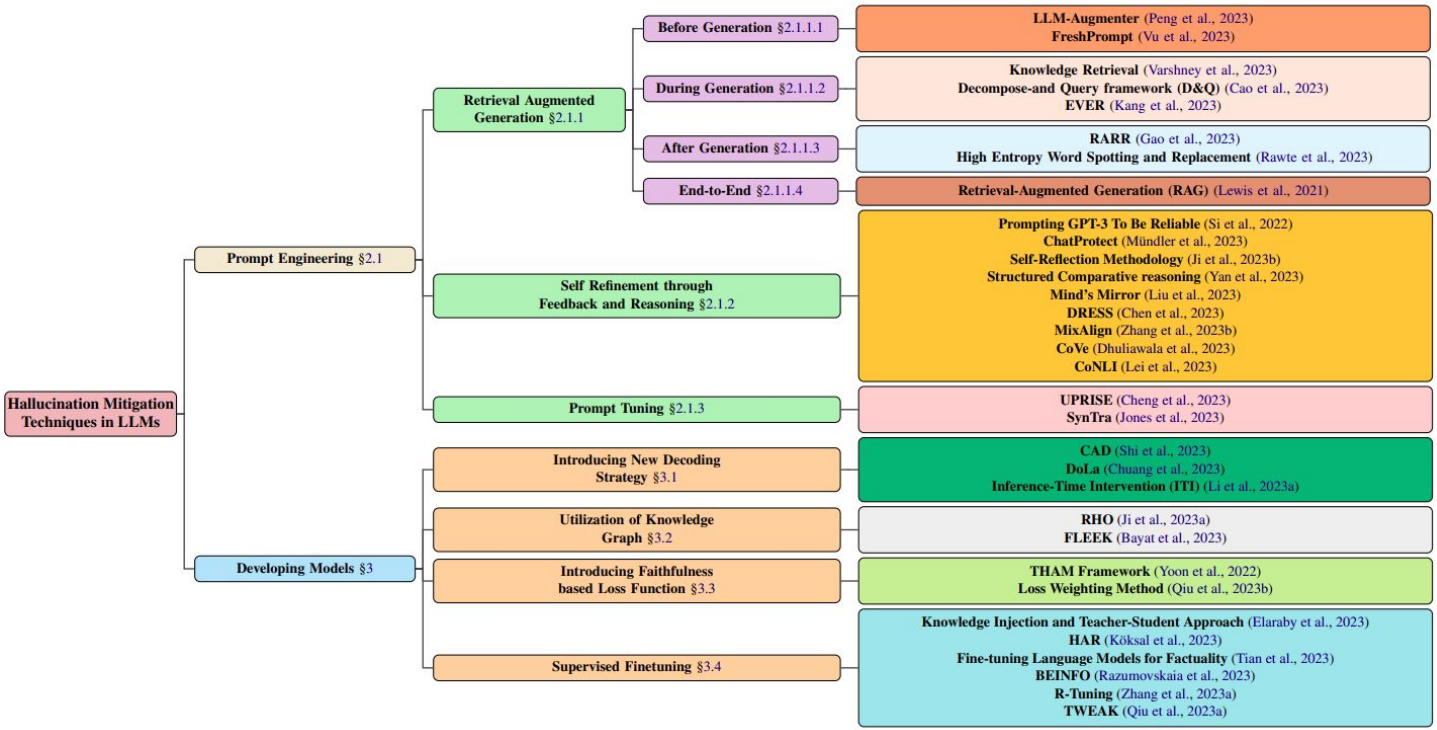
The only thing crazier than a guy in….tl:dr

Self-Supervised Training



Potential Harms of LLMs

- **Hallucination**



Potential Harms of LLMs

- **Toxic Language Generation**
  - meticulous data cleaning and filtering

- **Privacy Issues**
  - remove private datasets from the training data

Language models should include **datasheets** or **model cards**, giving full replicable information on the corpora used to train them.

Requirements that models are **transparent** in such ways is also in the process of being incorporated into the regulations of various national governments