

CX4240 Spring 2026

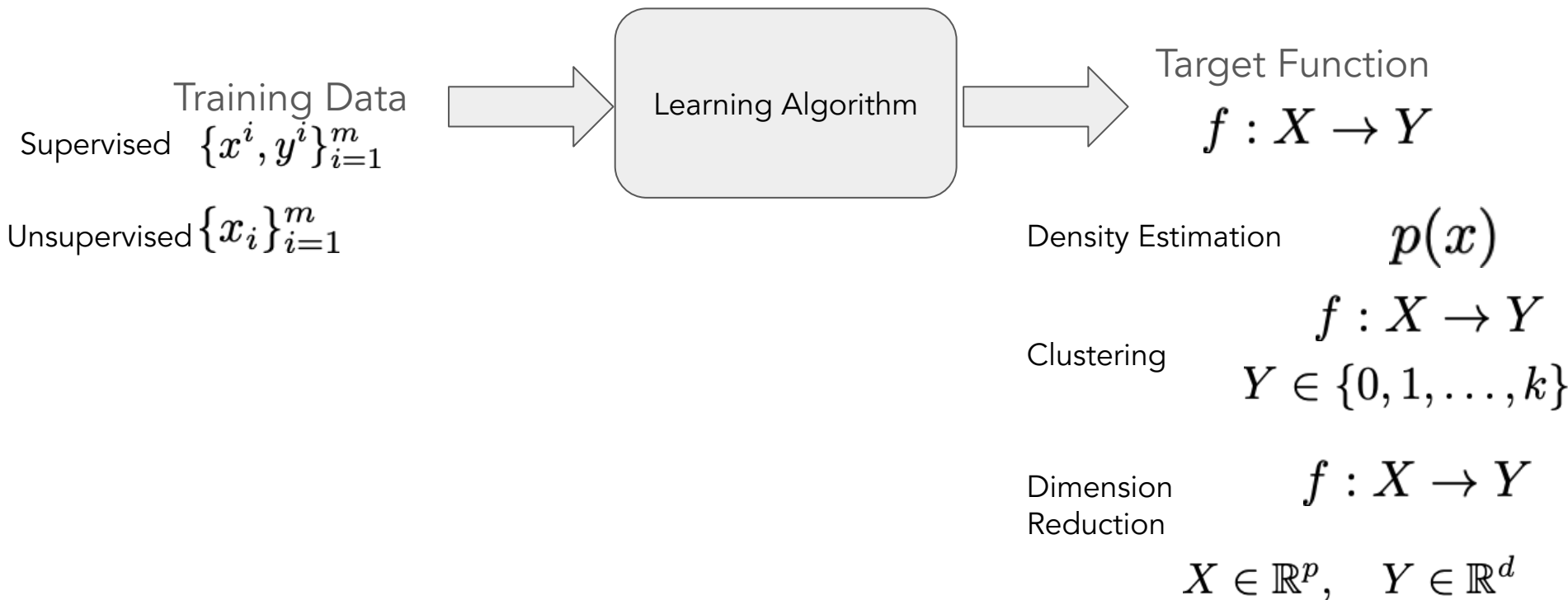
# (Large) Language Model (Part I): Attention and Transformers

Bo Dai  
School of CSE, Georgia Tech  
[bodai@cc.gatech.edu](mailto:bodai@cc.gatech.edu)

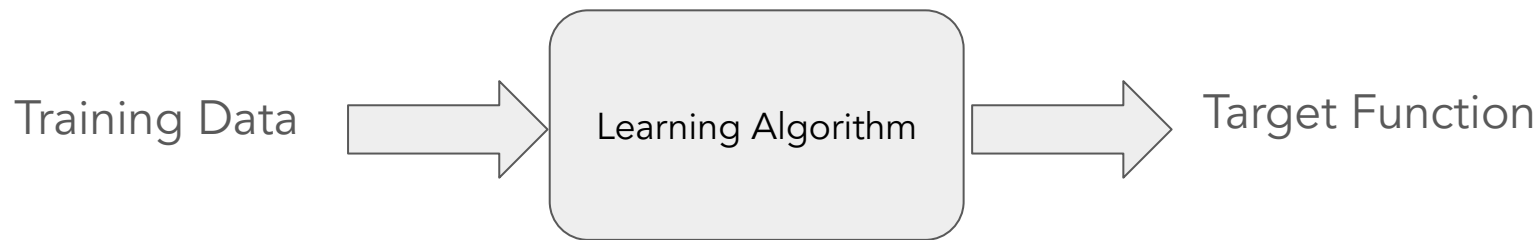
# Presentation

- Please [signup](#) the presentation date in group
  - Today is the **deadline!**
  - Group 7, 11, 18, and 21
- Final presentation will be online.
- Make sure to complete in time to get the bonus
  - Think this is a conference talk
- Final report is due on **May 4th.**

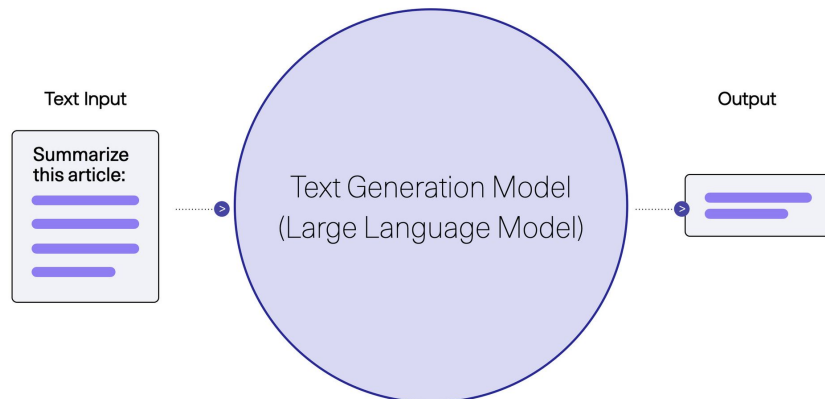
# Supervised Learning vs. Unsupervised Learning



# (Large) Language Models

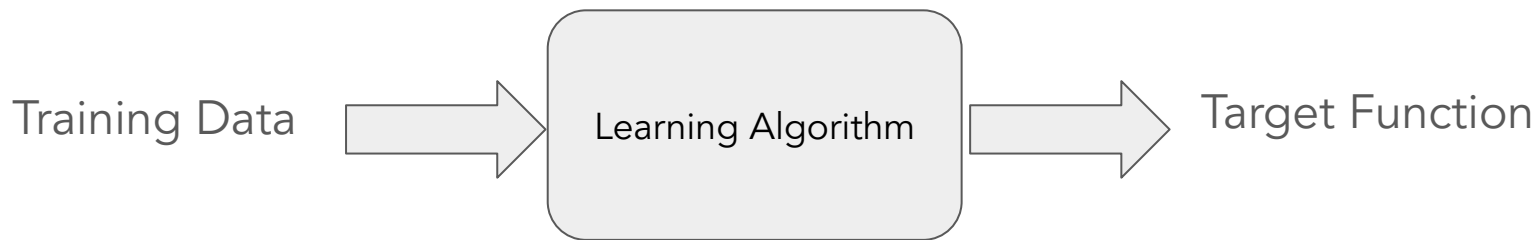


all text data

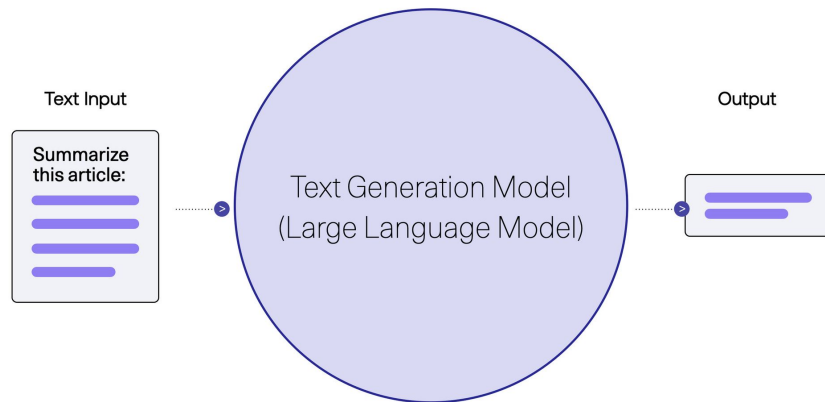


$$p(x) = \prod_{i=1}^k p(x_i | x_{<i})$$

# (Large) Language Models



all text data



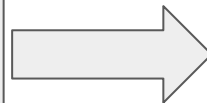
$$p(x_{l,\dots,k}|x_{<l}) = \prod_{i=l}^k p(x_i|x_{<i})$$

# (Large) Language Models

Training Data



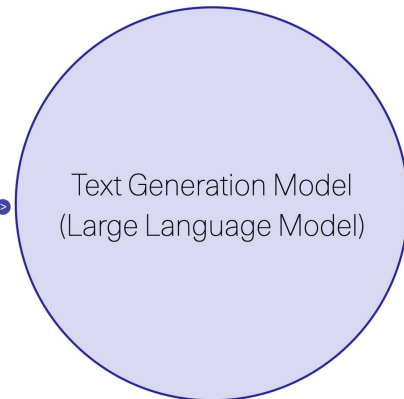
Learning Algorithm



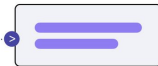
Target Function



Text Input



Output

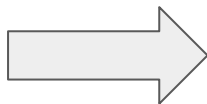


Unlabelled Data: text sequences

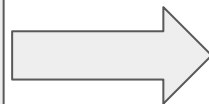
$$p(x) = \prod_{i=1}^k p(x_i | x_{<i})$$

# (Large) Language Models

Training Data



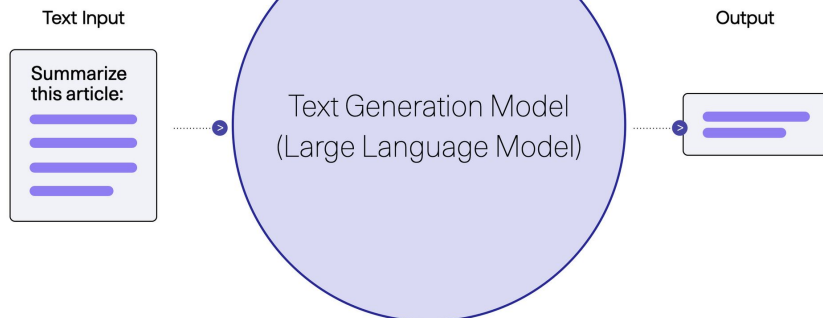
Learning Algorithm



Target Function



Unsupervised Learning



Unlabelled Data: text sequences

$$p(x) = \prod_{i=1}^k p(x_i | x_{<i})$$

# (Large) Language Models

Training Data



Learning Algorithm



Target Function

**Outcome (150-175 words):**  
Talk about your changing view on this belief or idea after the event you discussed in the Catalyst section (this is the 'after' picture you discussed in the Context section). If your old belief/idea was replaced with a new belief/idea, briefly explain what the new one is here.

Of course, avoidance doesn't fix anything. I had to face the music eventually and I have with time become more financially competent, spending hours researching and learning as much as I can. I've started an education savings plan and a long term savings account. I've become better at spending my money; I no longer blow my money on things I don't need. I will spend my pay and exactly how much is a know how to do yet. I know as I get older they can't even anticipate. I want to become better want to diversify. I don't really know what dive podcasts about finances and I read books. An money.

**Reflection (175-200 words):**  
Talk about what you learned by challenging this belief forward. Provide some learning outcomes about it professionally.

Every adult has to deal with money. I felt so gr a paycheck with my name on it. However, the believe it was a major shortcoming in my educ the hard way, slowly and with a lot of mistakes could have been better prepared prior if more money. I started out believing money was sor but I've been slowly shedding those practices people, and when I do I still feel like I'm break myself to speak plainly. And I have a goal, no going to pay what I learn forward. When my know what to do with it. I know it's important,

## MATH Dataset (Ours)

**Problem:** Tom has a red marble, a green marble, a blue marble, and three identical yellow marbles. How many different groups of two marbles can Tom choose?

**Solution:** There are two cases here: either Tom chooses two yellow marbles (1 result), or he chooses two marbles of different colors ( $\binom{4}{2} = 6$  results). The total number of distinct pairs of marbles Tom can choose is  $1 + 6 = 7$ .

**Problem:** If  $\sum_{n=0}^{\infty} \cos^{2n} \theta = 5$ , what is  $\cos 2\theta$ ?

**Solution:** This geometric series is  $1 + \cos^2 \theta + \cos^4 \theta + \dots = \frac{1}{1 - \cos^2 \theta} = 5$ . Hence,  $\cos^2 \theta = \frac{4}{5}$ . Then  $\cos 2\theta = 2 \cos^2 \theta - 1 = \frac{3}{5}$ .

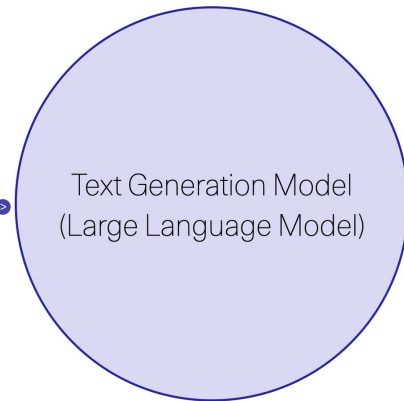
**Problem:** The equation  $x^2 + 2x = i$  has two complex solutions. Determine the product of their real parts.

**Solution:** Complete the square by adding 1 to each side. Then  $(x + 1)^2 = 1 + i = e^{\frac{\pi}{4}} \sqrt{2}$ , so  $x + 1 = \pm e^{\frac{\pi}{8}} \sqrt{2}$ . The desired product is then

$$\begin{aligned} & (-1 + \cos(\frac{\pi}{8}) \sqrt{2}) (-1 - \cos(\frac{\pi}{8}) \sqrt{2}) \\ &= 1 - \cos^2(\frac{\pi}{8}) \sqrt{2} = 1 - \frac{(1 + \cos(\frac{\pi}{4}))}{2} \sqrt{2} = \frac{1 - \sqrt{2}}{2} \end{aligned}$$

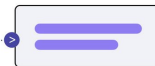
Text Input

Summarize this article:



Text Generation Model (Large Language Model)

Output



Labelled Data: Prompt & Answer

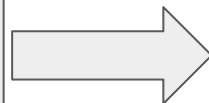
$$p(x_{l,\dots,k} | x_{<l}) = \prod_{i=l}^k p(x_i | x_{<i})$$

# (Large) Language Models

Training Data



Learning Algorithm



Target Function

**Outcome (150-175 words):**

Talk about your changing view on this belief or idea after the event you discussed in the Catalyst section (this is the 'after' picture you discussed in the Context section). If your old belief/idea was replaced with a new belief/idea, briefly explain what the new one is here.

Of course, avoidance doesn't fix anything. I had to face the music eventually and I have with time become more financially competent, spending hours researching and learning as much as I can. I've started an education savings plan and a long term savings account. I've become better at spending my money; I no longer blow money on things I don't need. I will spend my pay and exactly how much is a know how to do yet. I know as I get older they can't even anticipate. I want to become better want to diversify. I don't really know what 'dive podcasts about finances and I read books. An money.

**Reflection (175-200 words):**

Talk about what you learned by challenging this belief forward. Provide some learning outcomes about it professionally.

Every adult has to deal with money. I felt so guilty a paycheck with my name on it. However, the believe it was a major shortcoming in my educ the hard way, slowly and with a lot of mistakes could have been better prepared prior if more money. I started out believing money was something but I've been slowly shedding those practices people, and when I do I still feel like I'm breaking myself to speak plainly. And I have a goal, no solution. Complete the square by adding 1 to each side. Then  $(x+1)^2 = 1 + i = e^{\frac{\pi}{2}}\sqrt{2}$ , so  $x+1 = \pm e^{\frac{\pi}{4}}\sqrt{2}$ . The desired product is then  $(-1 + \cos(\frac{\pi}{8})\sqrt{2})(-1 - \cos(\frac{\pi}{8})\sqrt{2}) = 1 - \cos^2(\frac{\pi}{8})\sqrt{2} = 1 - \frac{(1+\cos(\frac{\pi}{4}))}{2}\sqrt{2} = \frac{1-\sqrt{2}}{2}$ .

**MATH Dataset (Ours)**

**Problem:** Tom has a red marble, a green marble, a blue marble, and three identical yellow marbles. How many different groups of two marbles can Tom choose?

**Solution:** There are two cases here: either Tom chooses two yellow marbles (1 result), or he chooses two marbles of different colors ( $\binom{4}{2} = 6$  results). The total number of distinct pairs of marbles Tom can choose is  $1 + 6 = 7$ .

**Problem:** If  $\sum_{n=0}^{\infty} \cos^{2n} \theta = 5$ , what is  $\cos 2\theta$ ?

**Solution:** This geometric series is  $1 + \cos^2 \theta + \cos^4 \theta + \dots = \frac{1}{1 - \cos^2 \theta} = 5$ . Hence,  $\cos^2 \theta = \frac{4}{5}$ . Then  $\cos 2\theta = 2 \cos^2 \theta - 1 = \frac{3}{5}$ .

**Problem:** The equation  $x^2 + 2x = i$  has two complex solutions. Determine the product of their real parts.

**Solution:** Complete the square by adding 1 to each side. Then  $(x+1)^2 = 1 + i = e^{\frac{\pi}{2}}\sqrt{2}$ , so  $x+1 = \pm e^{\frac{\pi}{4}}\sqrt{2}$ . The desired product is then

$$(-1 + \cos(\frac{\pi}{8})\sqrt{2})(-1 - \cos(\frac{\pi}{8})\sqrt{2}) = 1 - \cos^2(\frac{\pi}{8})\sqrt{2} = 1 - \frac{(1+\cos(\frac{\pi}{4}))}{2}\sqrt{2} = \frac{1-\sqrt{2}}{2}$$

## Supervised Learning

Text Input

Summarize this article:



Text Generation Model (Large Language Model)

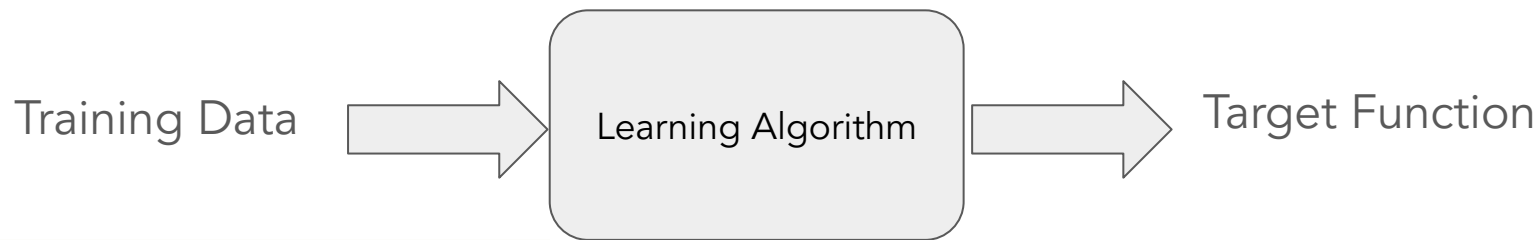
Output



Labelled Data: Prompt & Perfect Answer

$$p(x_{l,\dots,k} | x_{<l}) = \prod_{i=l}^k p(x_i | x_{<i})$$

# (Large) Language Models

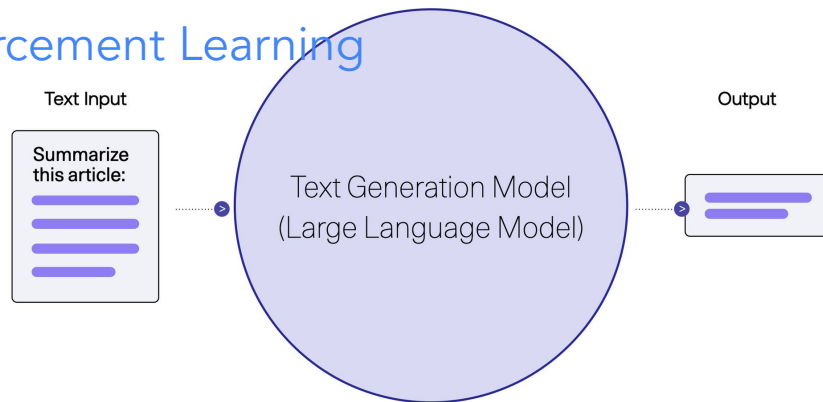


What are the key benefits of using Reinforcement Learning from Human Feedback (RLHF) for dataset collection in the context of Large Language Model (LLM) generation?

Reinforcement Learning from Human Feedback (RLHF) for dataset collection in Large Language Model (LLM) generation provides key benefits such as improved model performance through direct optimization, better alignment with human values by incorporating human feedback, and the ability to iteratively refine the model based on user interactions, resulting in a more user-friendly and efficient language model.

Using reinforcement Learning from Human Feedback (RLHF) for dataset collection in Large Language Model (LLM) generation offers advantages such as enhanced model capabilities by optimizing for desired outcomes, greater adaptability to human preferences through the inclusion of human feedback, and the opportunity to continuously improve the model based on user experiences, ultimately leading to a more effective and responsive language model.

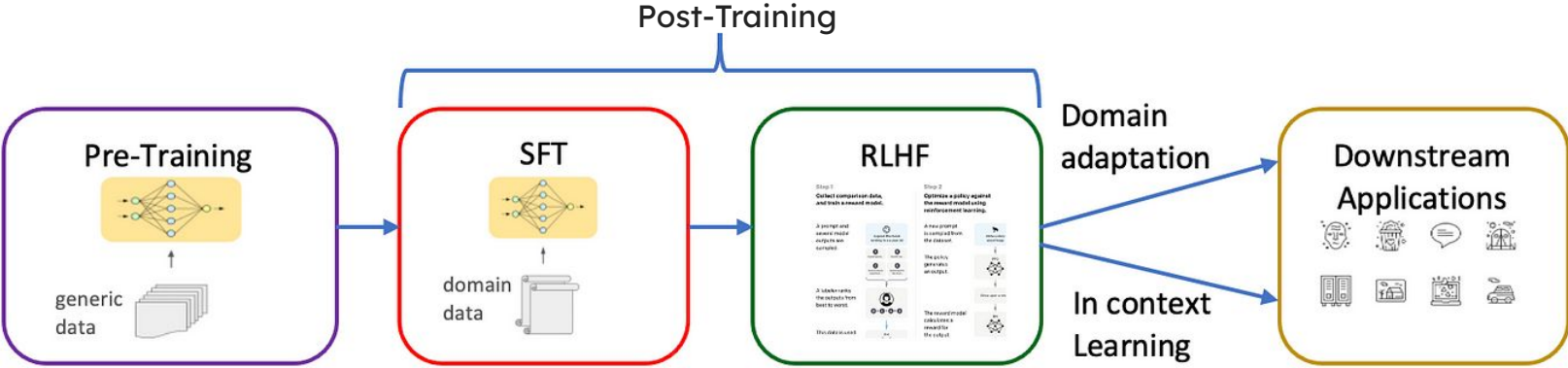
## Reinforcement Learning



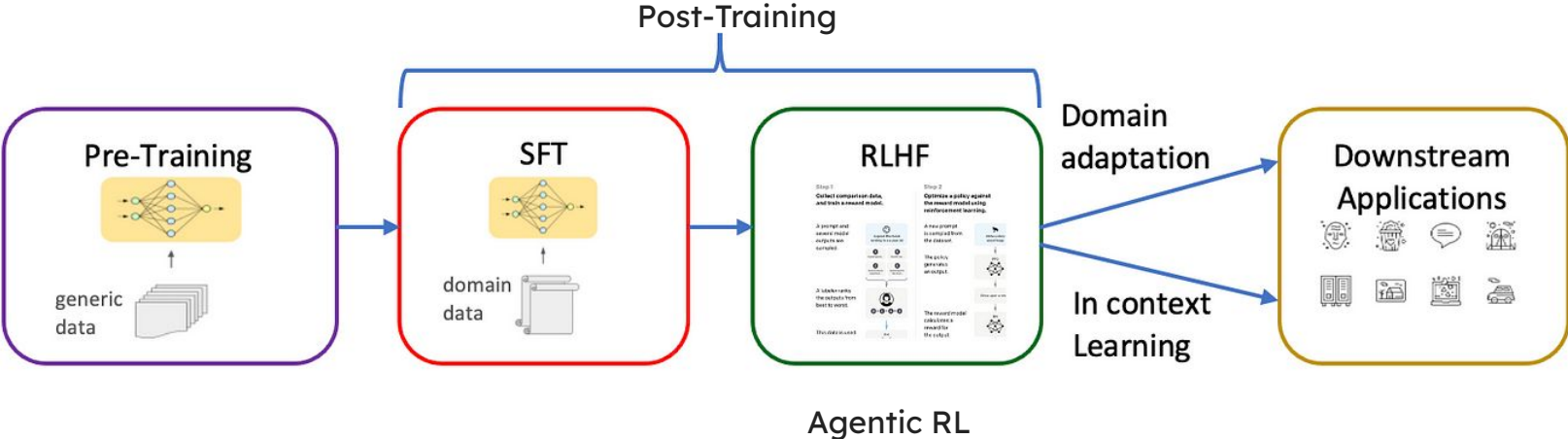
Preference Data:  
Prompt & Positive/ Negative Answers

$$p(x_{l,\dots,k}|x_{<l}) = \prod_{i=l}^k p(x_i|x_{<i})$$

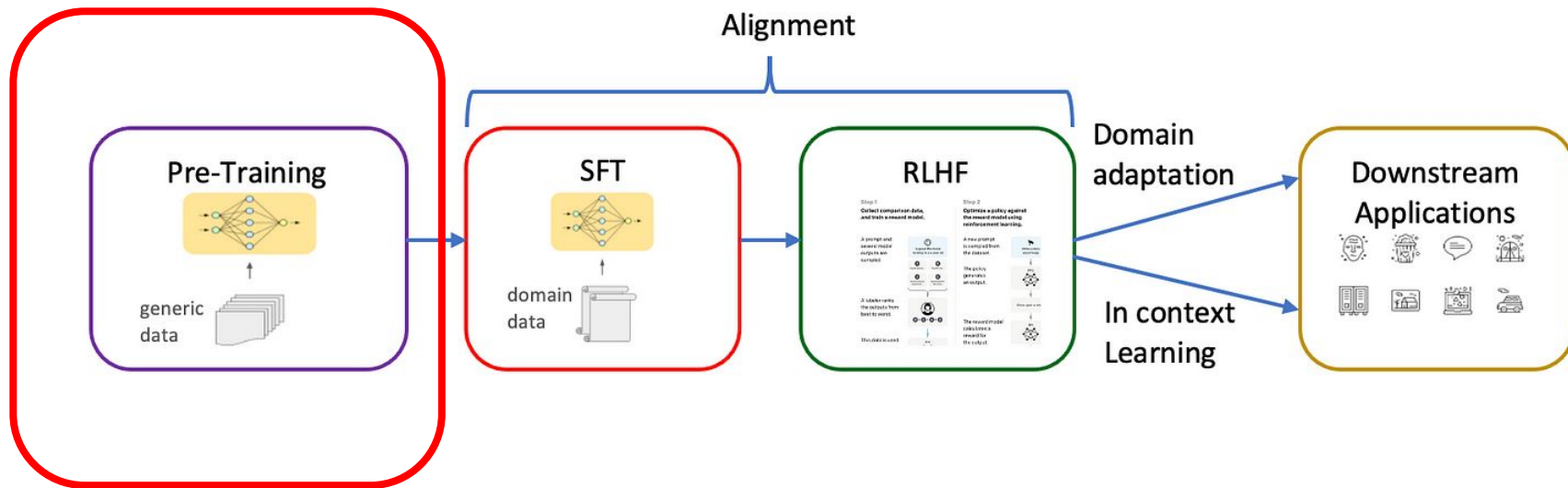
# LLM Training



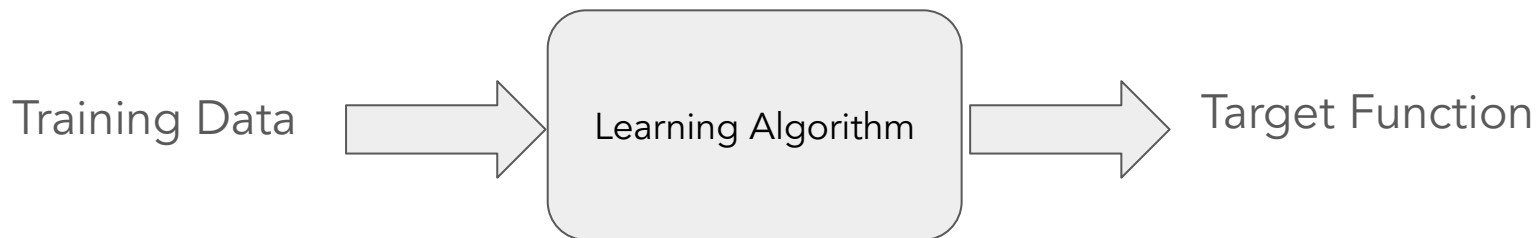
# LLM Training



# LLM Training



# (Large) Language Models

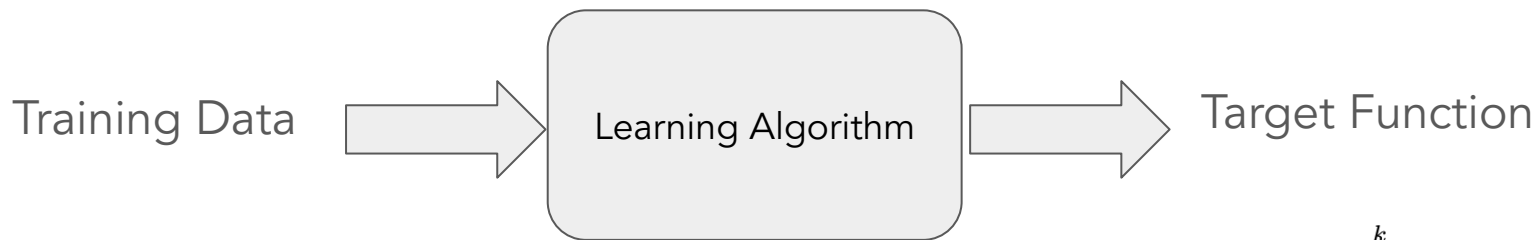


For a movie that gets no respect there sure are a lot of memorable quotes listed for this gem. Imagine a movie where Joe Piscopo is actually funny! Maureen Stapleton is a scene stealer. The Moroni character is an absolute scream. Watch for Alan "The Skipper" Hale jr. as a police Sgt.

$p(X)$

$$D = \{X^i\}_{i=1}^n, \quad X^i = \{x_j^i\}_{j=1}^t$$

# (Large) Language Models



$$D = \{X^i\}_{i=1}^n, \quad X^i = \{x_j^i\}_{j=1}^t$$

$$p(x) = \prod_{i=1}^k p(x_i | x_{<i})$$

1. Build probabilistic models  
Categorical Distribution + Autoregressive +  
RNN->Transformer
2. Derive loss function (by MLE or MAP....)
3. Select optimizer

# Combating Combinatorial Complexity: Autoregressive Model

$$p(X) = p(\{x_j\}_{j=1}^t) \quad O(|V|^T)$$

# Combating Combinatorial Complexity: Autoregressive Model

$$\begin{aligned} p(X) &= p(\{x_j\}_{j=1}^t) && O(|V|^T) \\ &= \prod_{j=1}^t p(x_j | x_{<j}) \end{aligned}$$

# Categorical Distribution

$$\begin{aligned} p(X) &= p(\{x_j\}_{j=1}^t) && O(|V|^T) \\ &= \prod_{j=1}^t p(x_j | x_{<j}) \\ &= \prod_{j=1}^t \frac{\exp(W_{x_j} \phi(x_{<j}))}{\sum_{l=1}^V \exp(W_l \phi(x_{<j}))} \end{aligned}$$

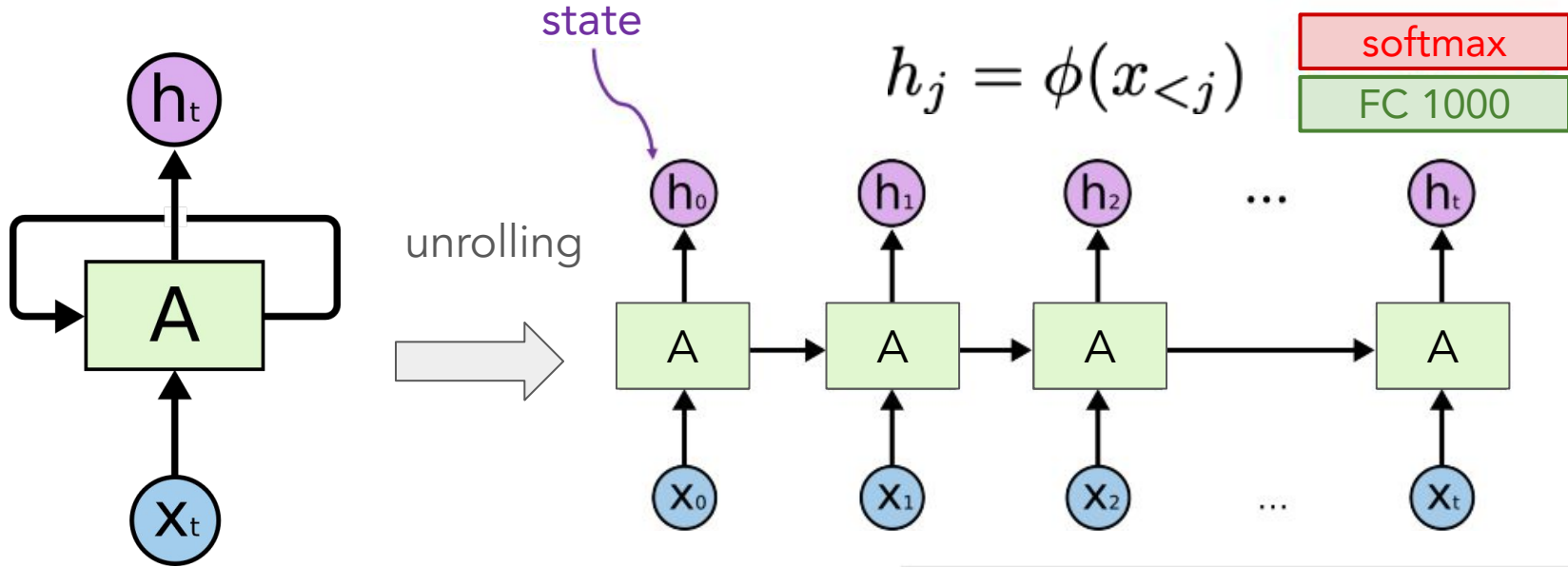
# Recursive Neural Network

$$\begin{aligned} p(X) &= p(\{x_j\}_{j=1}^t) && O(|V|^T) \\ &= \prod_{j=1}^t p(x_j | x_{<j}) \\ &= \prod_{j=1}^t \frac{\exp(W_{x_j} \phi(x_{<j}))}{\sum_{l=1}^V \exp(W_l \phi(x_{<j}))} \end{aligned}$$

RNN

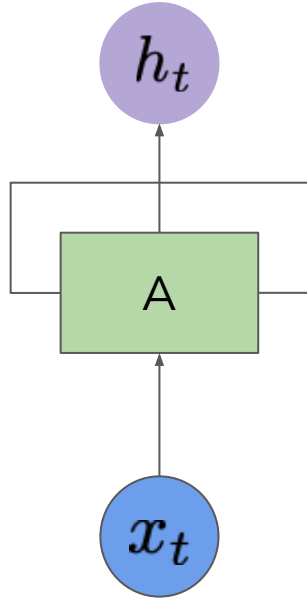
# Recurrent Neural Network

$$\frac{\exp(W_{x_j} \phi(x_{<j}))}{\sum_{l=1}^V \exp(W_l \phi(x_{<j}))}$$

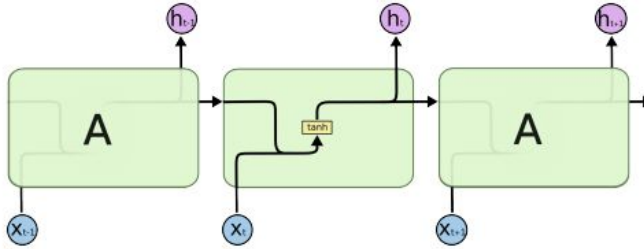


For a movie that gets no respect there sure are a lot of memorable quotes listed for this gem. Imagine a movie where Joe Piscopo is actually funny! Maureen Stapleton is a scene stealer. The Moroni character is an absolute scream. Watch for Alan "The Skipper" Hale jr. as a police Sgt.

# RNN Cell



$$\frac{\exp(W_{x_j} \phi(x_{<j}))}{\sum_{l=1}^V \exp(W_l \phi(x_{<j}))}$$

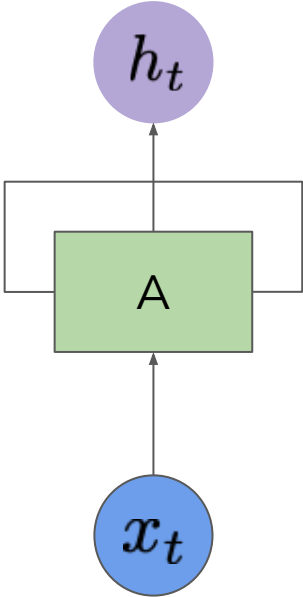


Simple RNN

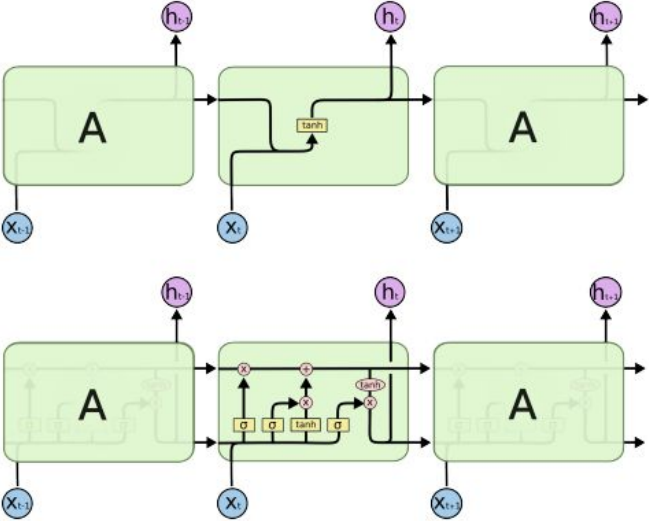
$\mathbf{h}_{100}$  is almost irrelevant to  $\mathbf{x}_1$ :  $\frac{\partial \mathbf{h}_{100}}{\partial \mathbf{x}_1}$  is near zero.

Gradient Vanishing

# RNN Cell



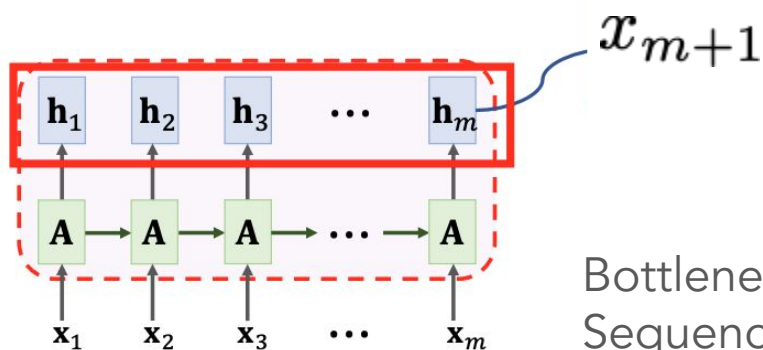
$$\frac{\exp(W_{x_j} \phi(x_{<j}))}{\sum_{l=1}^V \exp(W_l \phi(x_{<j}))}$$



Simple RNN

LSTM

# Bottleneck in RNN

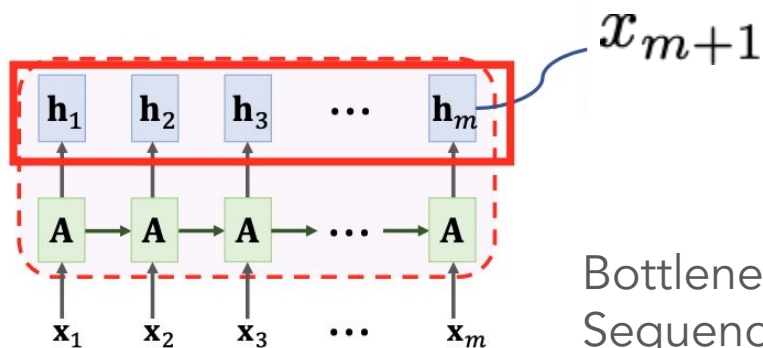


$$h_j = \phi(x_{<j})$$

$$\frac{\exp(W_{x_j} \phi(x_{<j}))}{\sum_{l=1}^V \exp(W_l \phi(x_{<j}))}$$

Bottleneck:  
Sequences bottlenecked through a  
fixed-sized vector.

# Bottleneck in RNN



$$h_j = \phi(x_{<j})$$

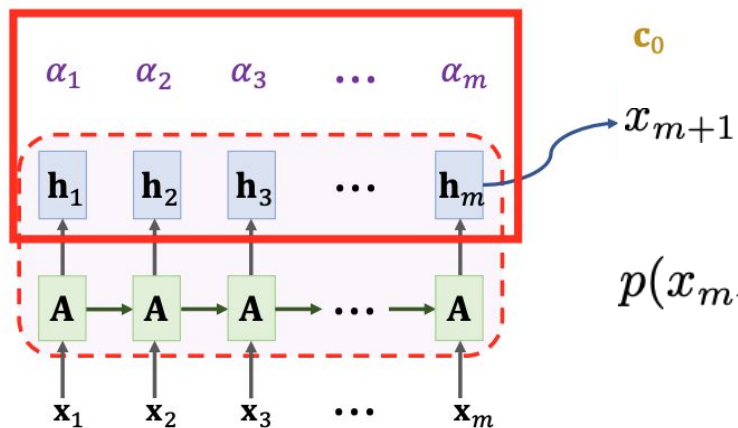
$$\frac{\exp(W_{x_j} \phi(x_{<j}))}{\sum_{l=1}^V \exp(W_l \phi(x_{<j}))}$$

Bottleneck:  
Sequences bottlenecked through a  
**fixed-sized** vector.

**Fixed size** -> **Length dependent**

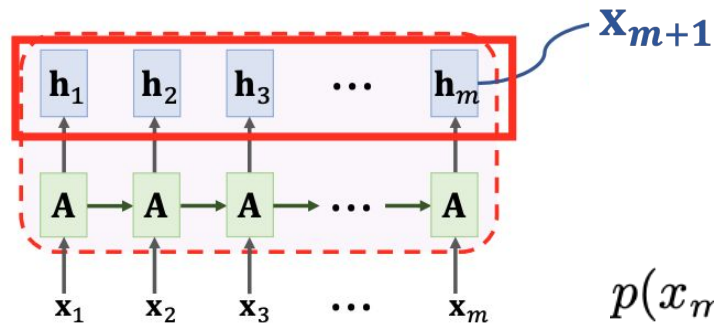
# Attention: Flatten RNN Computation

**Context vector:**  $\mathbf{c}_0 = \alpha_1 \mathbf{h}_1 + \dots + \alpha_m \mathbf{h}_m$ .



$$p(\mathbf{x}_{m+1} | \mathbf{x}_{<m+1}) = \frac{\exp(W_{x_j} \mathbf{c}_0)}{\sum_{l=1}^V \exp(W_l \mathbf{c}_0)}$$

# Attention: Adding Output Dependency

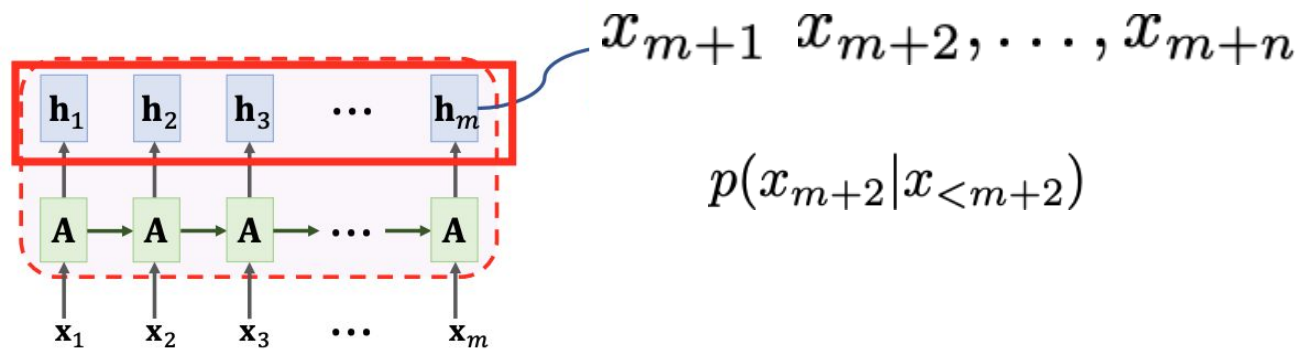


1. Linear maps:
  - $\mathbf{k}_i = \mathbf{W}_K \cdot \mathbf{h}_i$ , for  $i = 1$  to  $m$ .
  - $\mathbf{q}_0 = \mathbf{W}_Q \cdot \mathbf{x}_{m+1}$
2. Inner product:
  - $\tilde{\alpha}_i = \mathbf{k}_i^T \mathbf{q}_0$ , for  $i = 1$  to  $m$ .
3. Normalization:
  - $[\alpha_1, \dots, \alpha_m] = \text{Softmax}([\tilde{\alpha}_1, \dots, \tilde{\alpha}_m])$ .

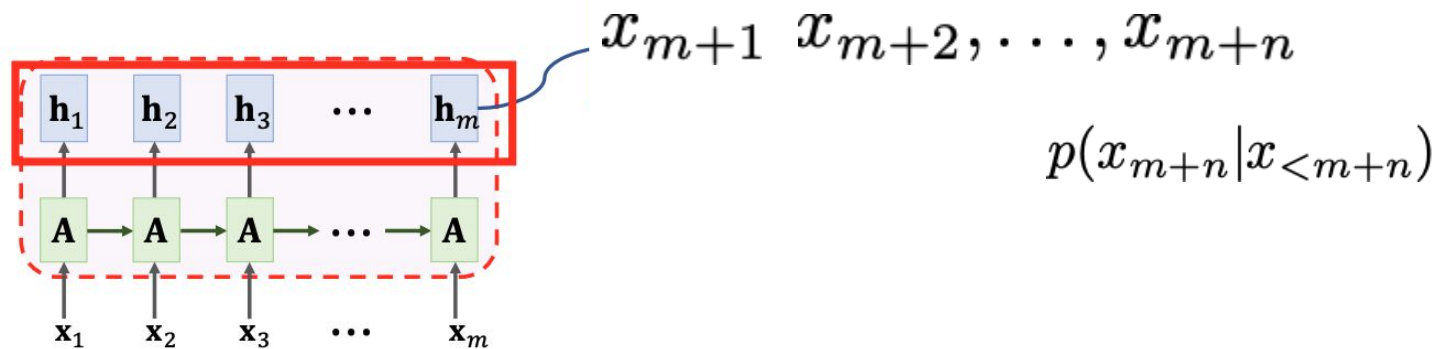
$$p(x_{m+1} | x_{<m+1}) = \frac{\exp(W_{x_j} c_0)}{\sum_{l=1}^V \exp(W_l c_0)}$$

**Weight:**  $\alpha_i = \text{align}(\mathbf{h}_i, \mathbf{x}_{m+1})$

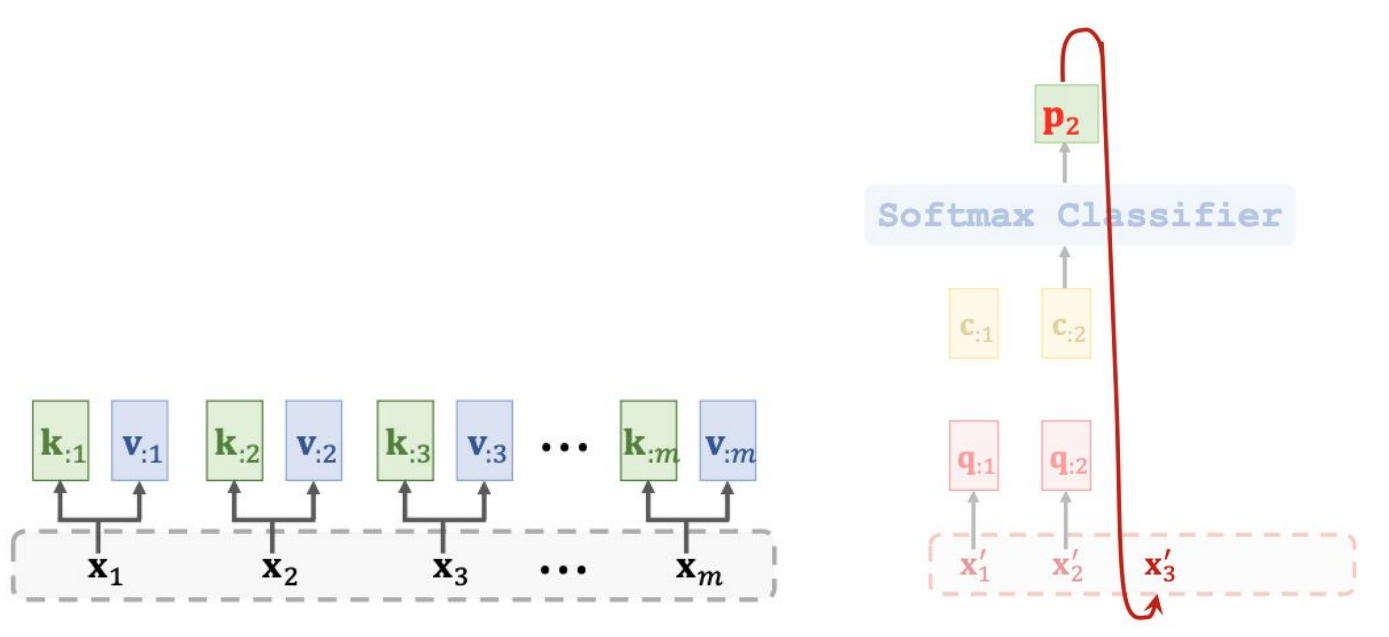
# AutoRecursive Attention



# AutoRecursive Attention

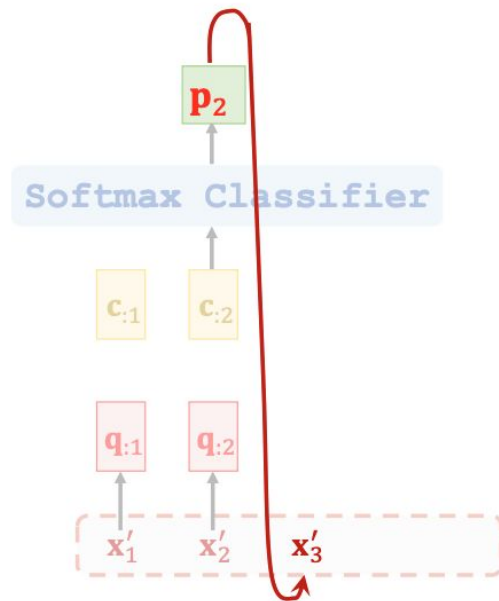
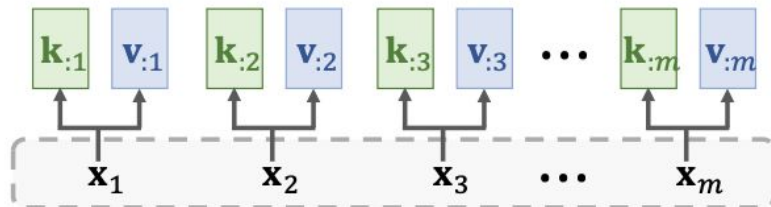


# AutoRecursive Attention for Inference

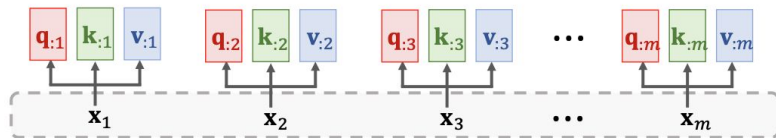


# AutoRecursive Attention for Inference

Not flexible enough

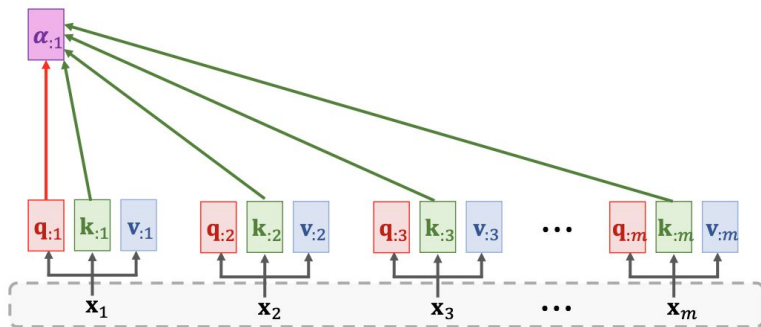


# Transformer: Multi-Layer Multi-Headed Self-Attention!



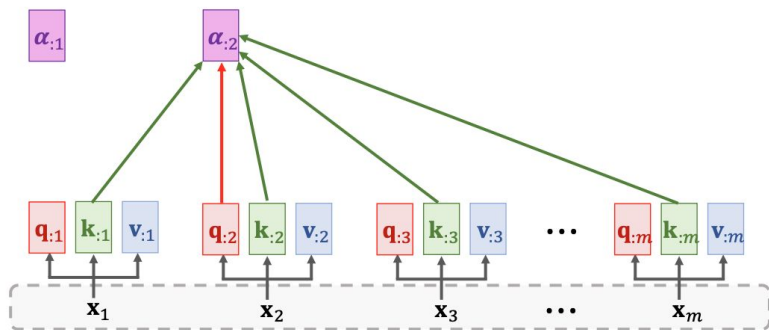
# Self-Attention

**Weights:**  $\alpha_j = \text{Softmax}(\mathbf{K}^T \mathbf{q}_j) \in \mathbb{R}^m$ .



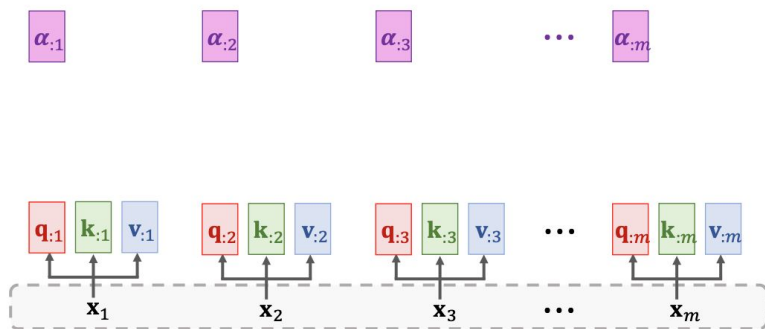
# Self-Attention

**Weights:**  $\alpha_j = \text{Softmax}(\mathbf{K}^T \mathbf{q}_j) \in \mathbb{R}^m$ .



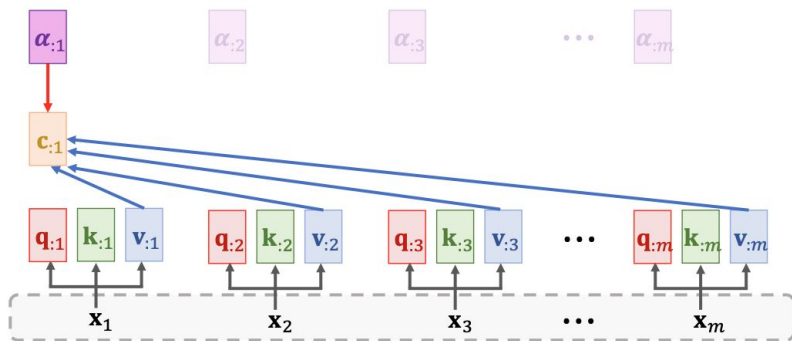
# Self-Attention

**Weights:**  $\alpha_{:j} = \text{Softmax}(\mathbf{K}^T \mathbf{q}_{:j}) \in \mathbb{R}^m.$



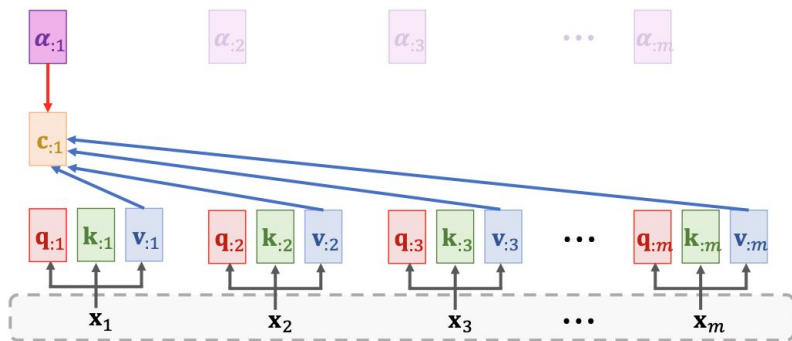
# Self-Attention

**Context vector:**  $\mathbf{c}_{:1} = \alpha_{11}\mathbf{v}_{:1} + \dots + \alpha_{m1}\mathbf{v}_{:m} = \mathbf{V}\alpha_{:1}$ .



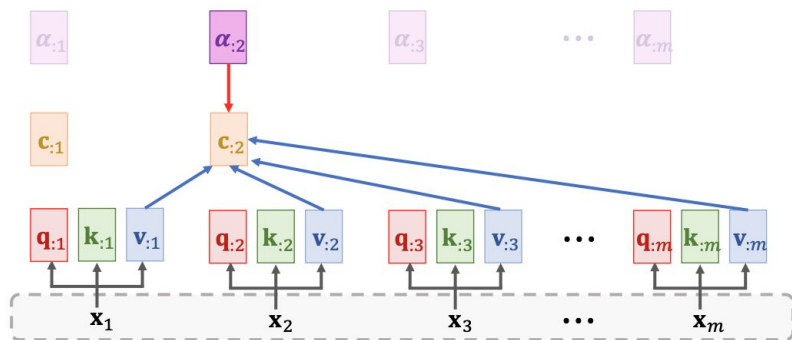
# Self-Attention

**Context vector:**  $\mathbf{c}_{:1} = \alpha_{11}\mathbf{v}_{:1} + \dots + \alpha_{m1}\mathbf{v}_{:m} = \mathbf{V}\alpha_{:1}$ .



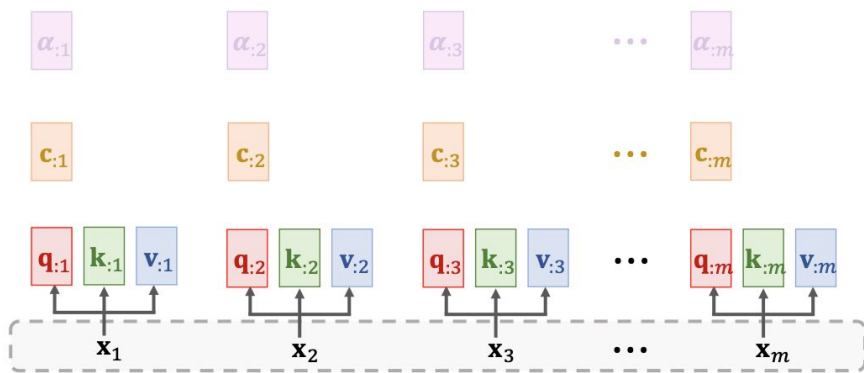
# Self-Attention

**Context vector:**  $\mathbf{c}_{:1} = \alpha_{11}\mathbf{v}_{:1} + \dots + \alpha_{m1}\mathbf{v}_{:m} = \mathbf{V}\boldsymbol{\alpha}_{:1}$ .



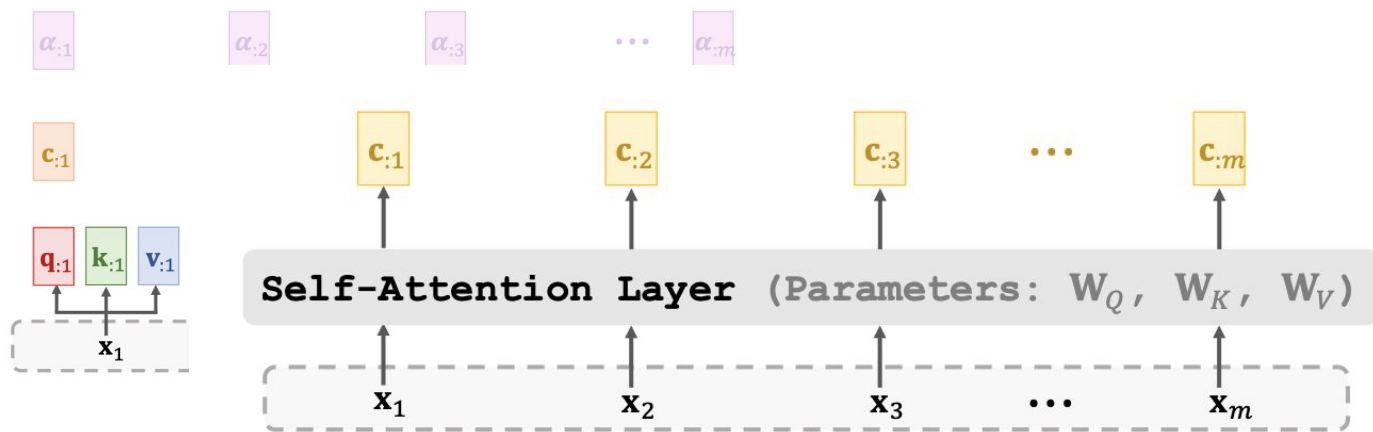
# Self-Attention

**Context vector:**  $\mathbf{c}_{:1} = \alpha_{11}\mathbf{v}_{:1} + \dots + \alpha_{m1}\mathbf{v}_{:m} = \mathbf{V}\boldsymbol{\alpha}_{:1}$ .

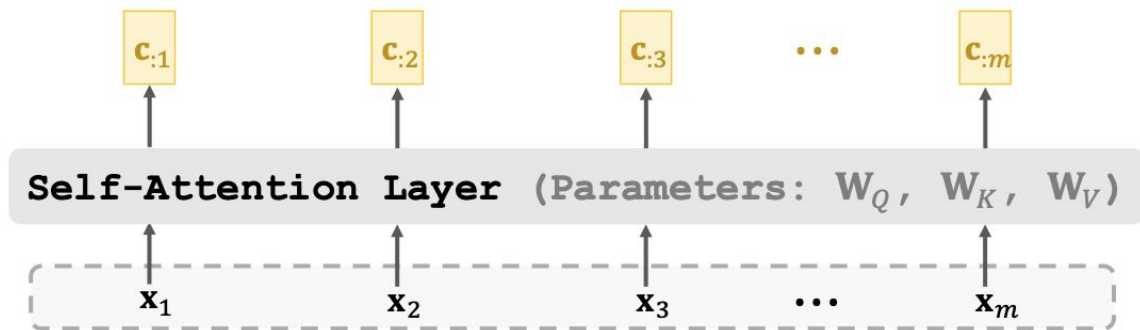


# Self-Attention

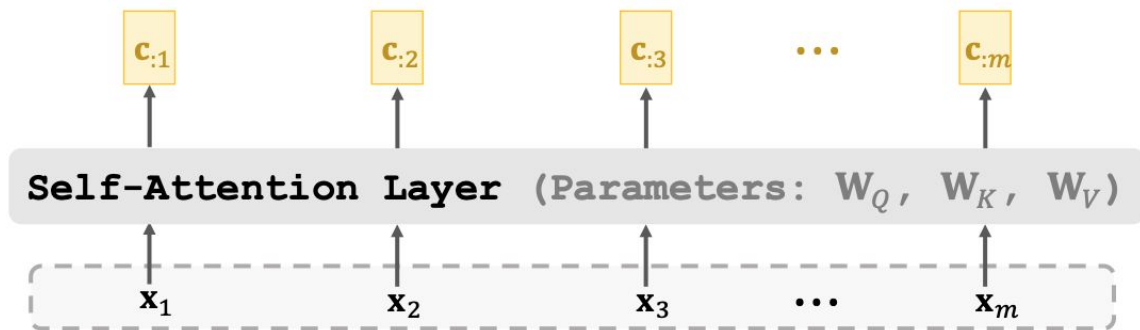
**Context vector:**  $\mathbf{c}_{:1} = \alpha_{11}\mathbf{v}_{:1} + \dots + \alpha_{m1}\mathbf{v}_{:m} = \mathbf{V}\alpha_{:1}$ .



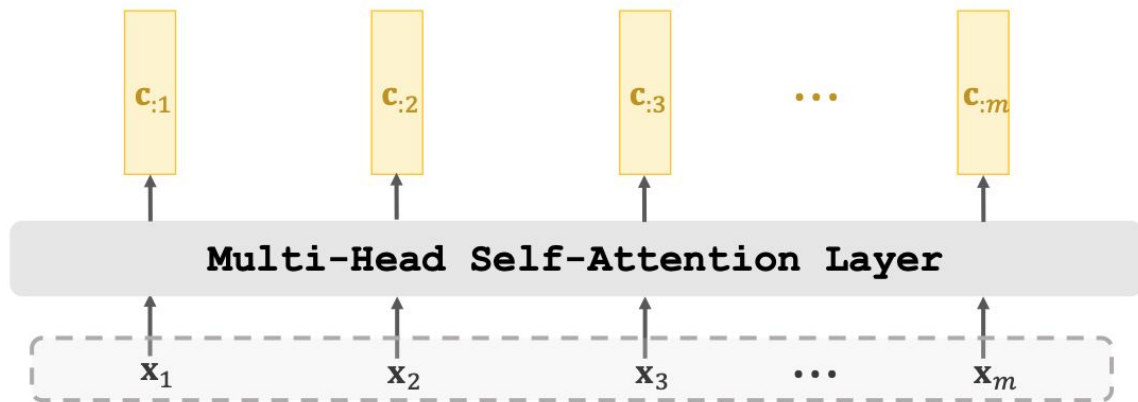
# Transformer: Multi-Layer Multi-Headed Self-Attention!



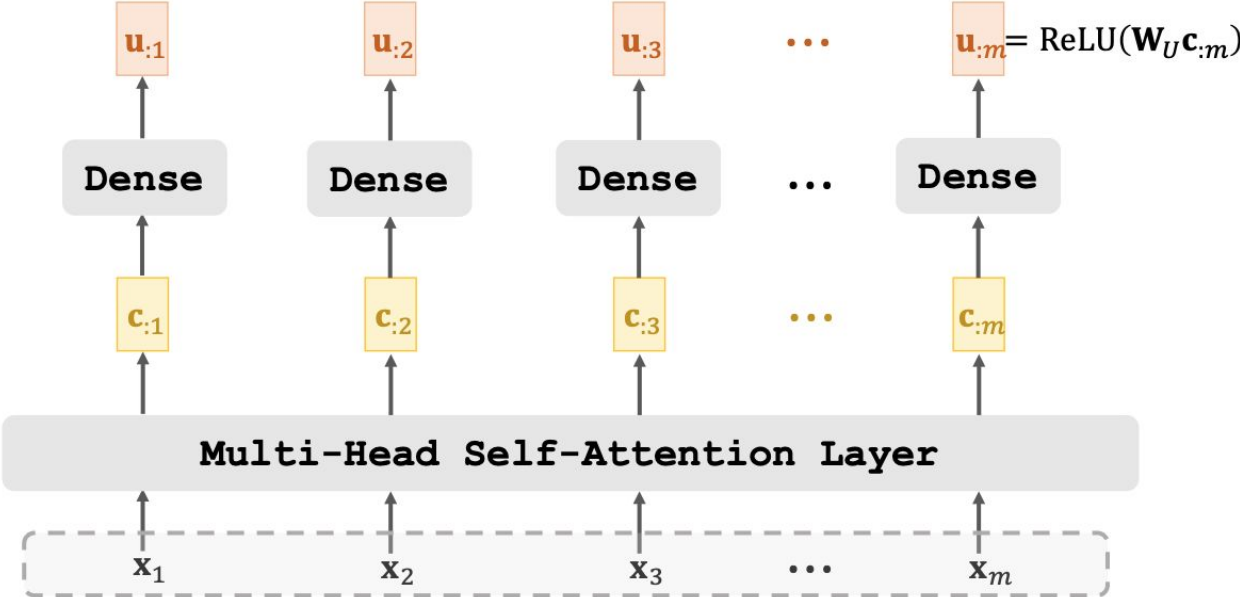
# Multi-Headed Self-Attention



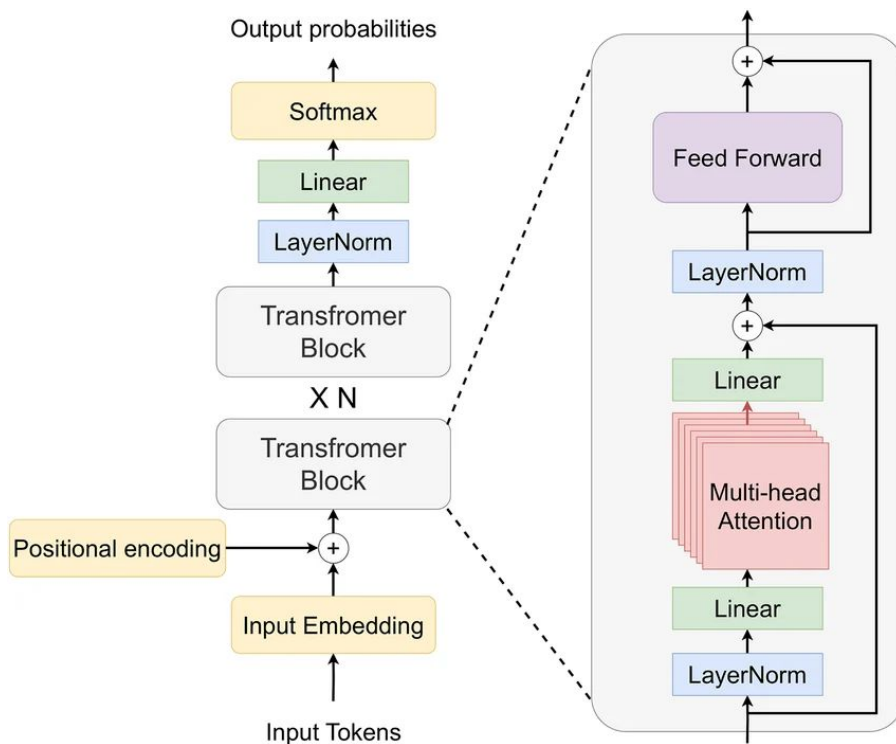
# Multi-Headed Self-Attention



# Multi-Headed Self-Attention



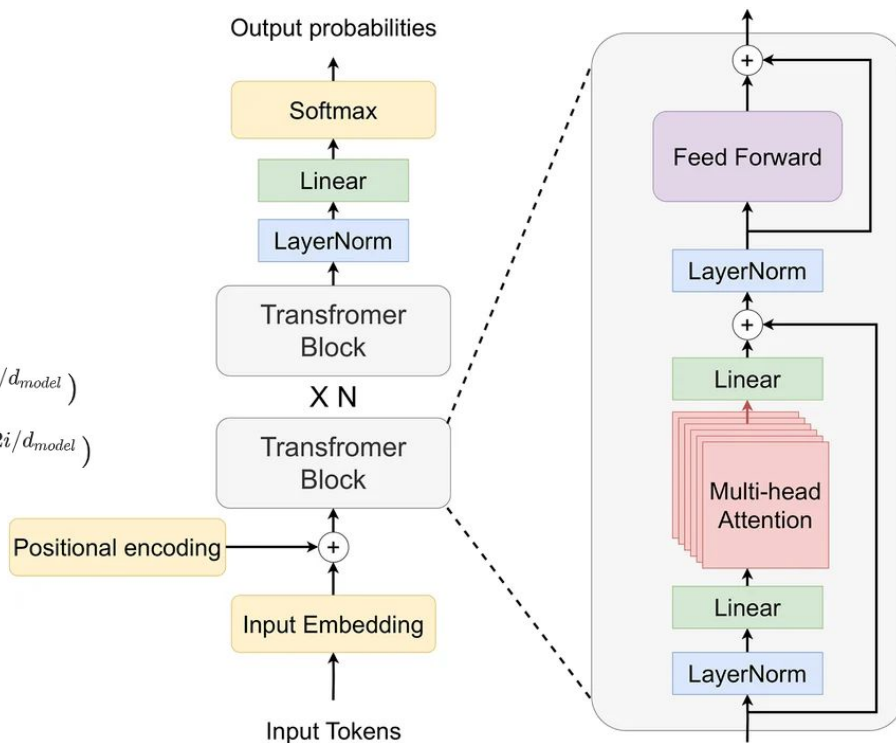
# Transformer: Multi-Layer Multi-Headed Self-Attention!



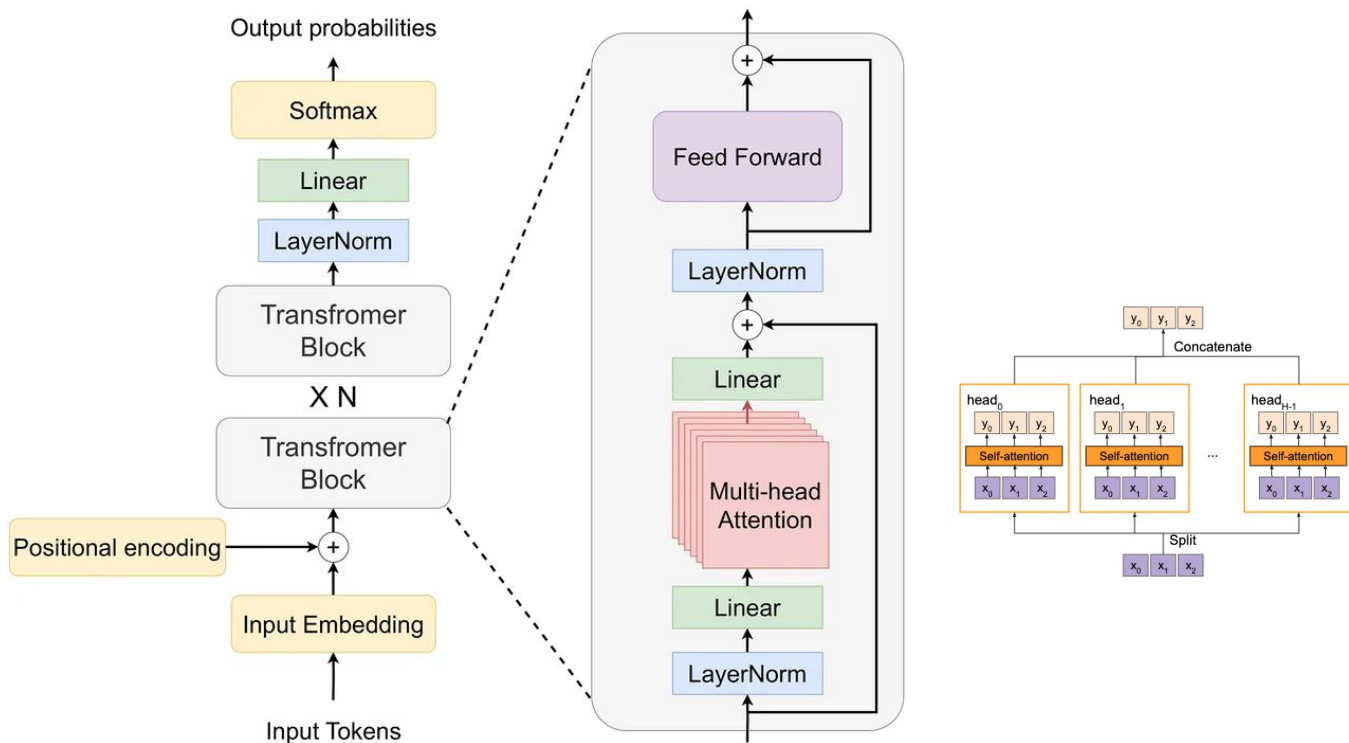
# Transformer: Multi-Layer Multi-Headed Self-Attention!

$$PE_{(pos,2i)} = \sin(pos/10000^{2i/d_{model}})$$

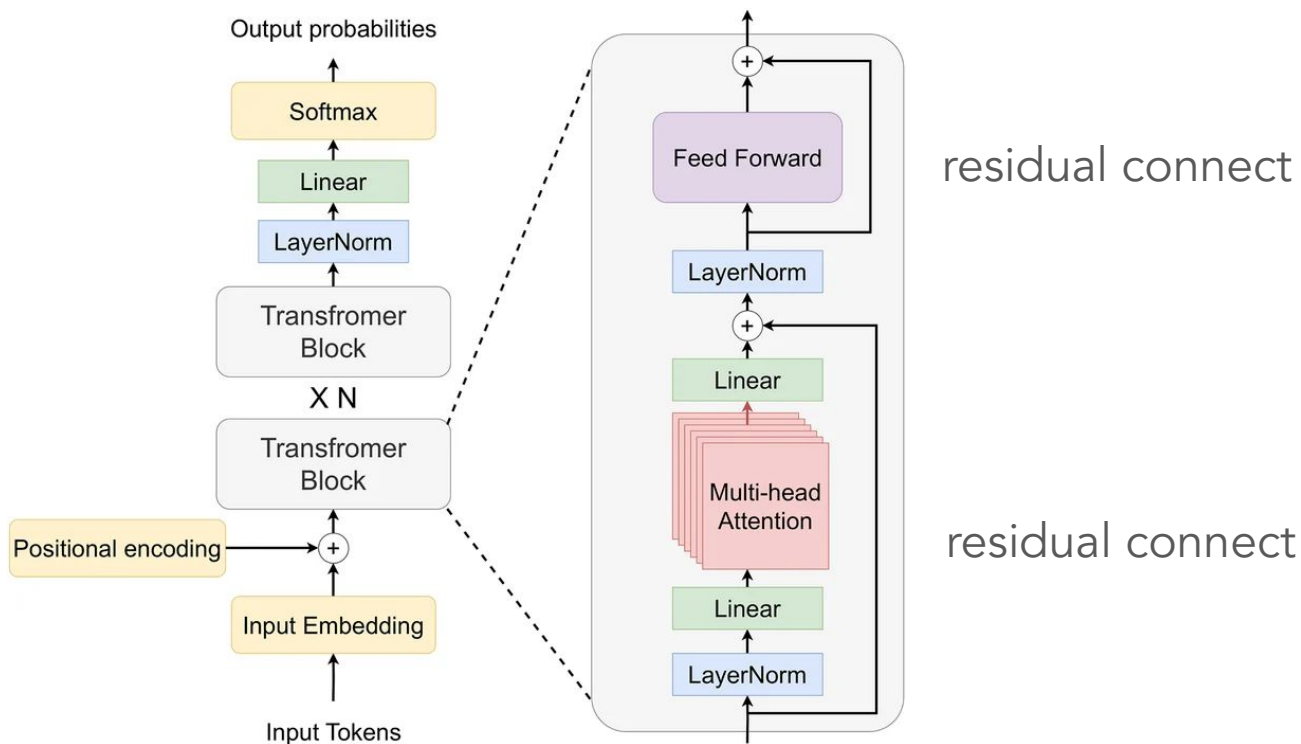
$$PE_{(pos,2i+1)} = \cos(pos/10000^{2i/d_{model}})$$



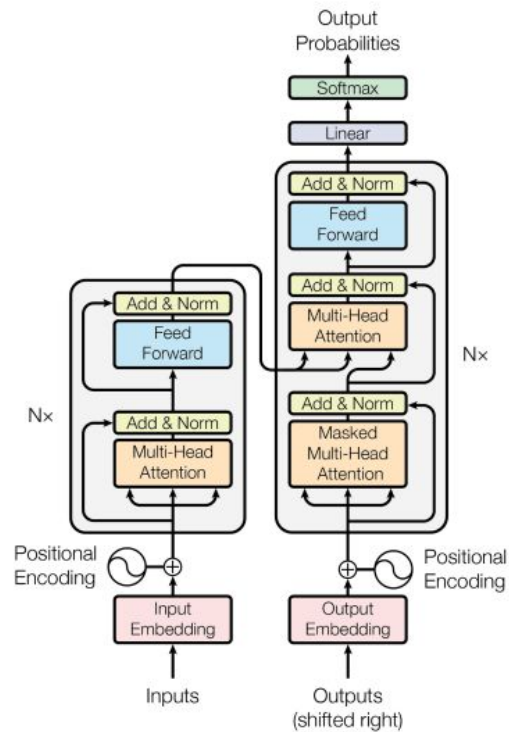
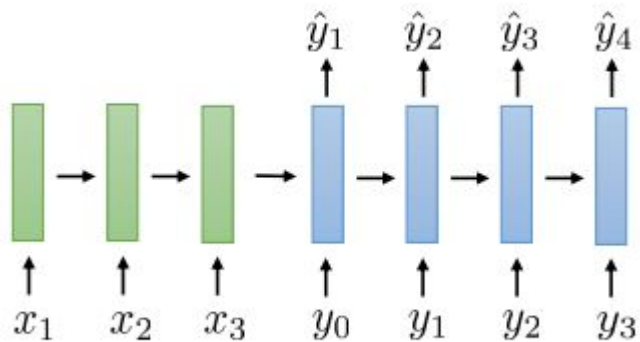
# Transformer: Multi-Layer Multi-Headed Self-Attention!



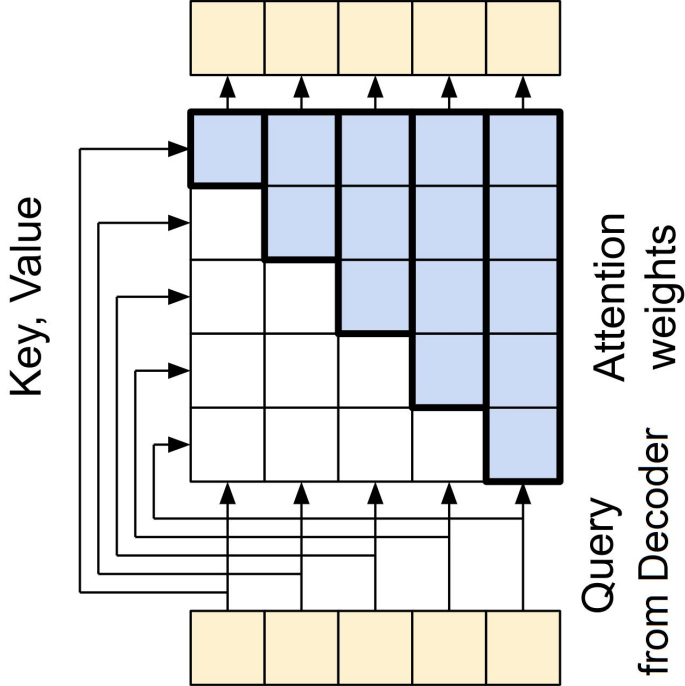
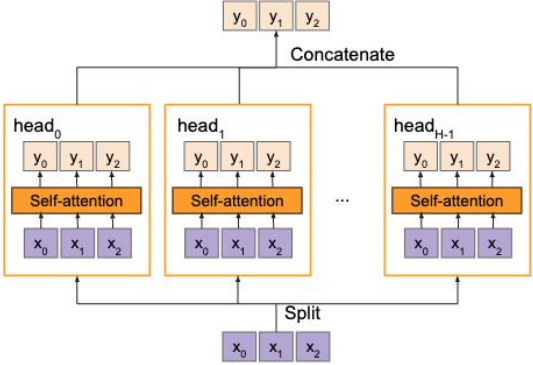
# Transformer: Multi-Layer Multi-Headed Self-Attention!



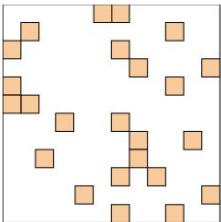
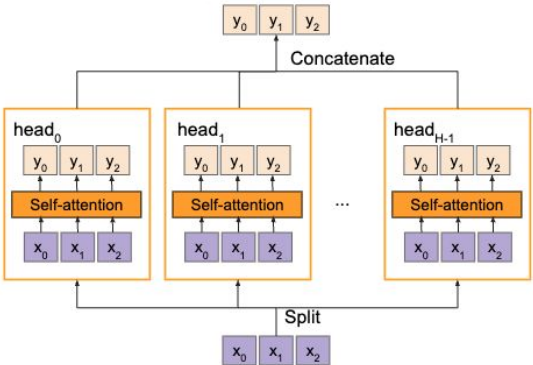
# Original Transformer



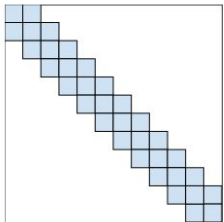
# Limitations? Expensive Computation



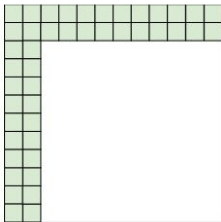
# Limitations?



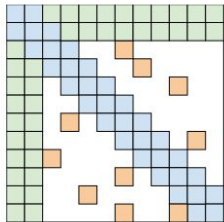
(a) Random attention



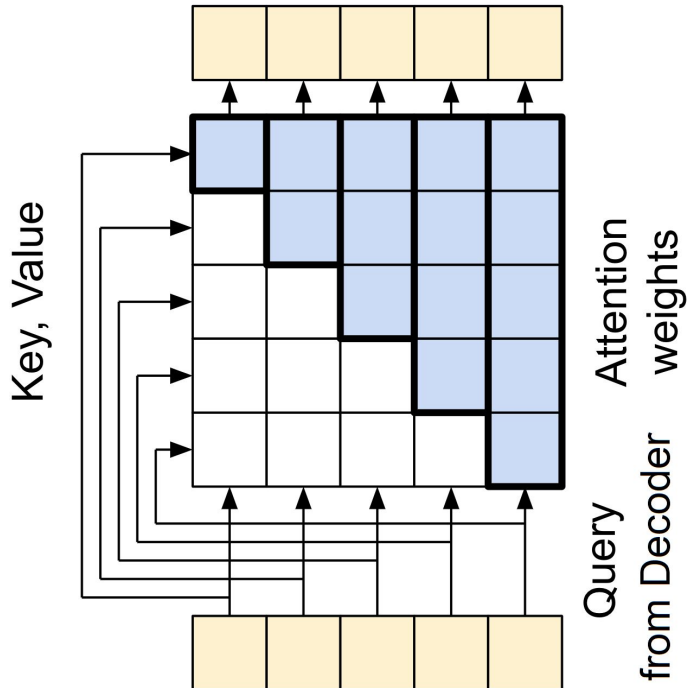
(b) Window attention



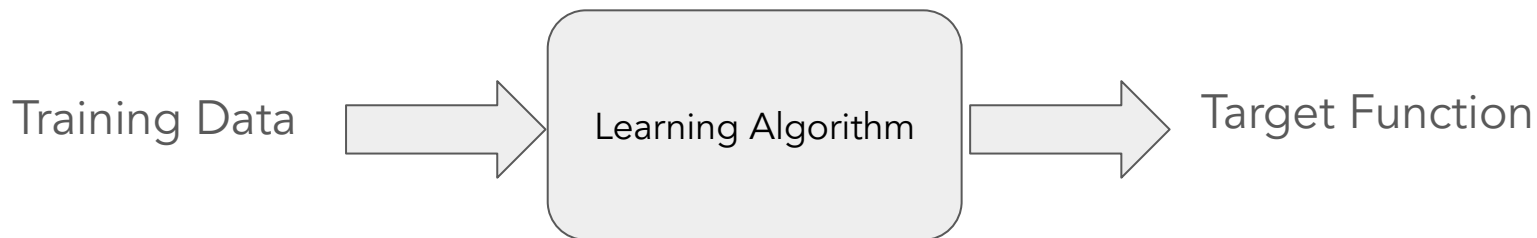
(c) Global Attention



(d) BIGBIRD



# (Large) Language Models



$$D = \{X^i\}_{i=1}^n, \quad X^i = \{x_j^i\}_{j=1}^t$$

$$p(X)$$

1. Build probabilistic models
2. Derive loss function (by MLE or MAP...)  
MLE
3. Select optimizer

# MLE

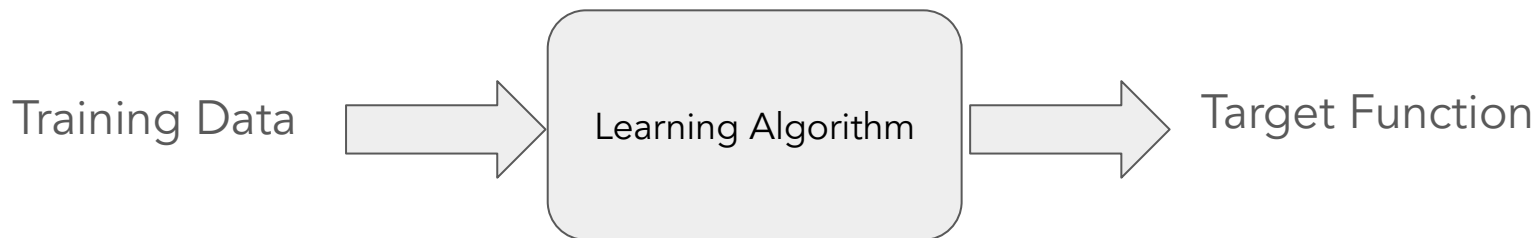
- Given all input data  $D = \{X^i\}_{i=1}^n$ ,  $X^i = \{x_j^i\}_{j=1}^t$

$$p(X^i) = \prod_{j=1}^t \frac{\exp(W_{x_j^i} \phi(x_{<j}^i))}{\sum_{l=1}^V \exp(W_l \phi(x_{<j}^i))}$$

- Log-likelihood

$$\begin{aligned} \ell(\phi) &= \sum_{i=1}^n p(X^i; \phi) = \sum_{i=1}^n \sum_{j=1}^t \log(x_j^i | x_{<j}^i; \phi) \\ &= \sum_{i=1}^n \sum_{j=1}^t \log \frac{\exp(W_{x_j^i} \phi(x_{<j}^i))}{\sum_{l=1}^V \exp(W_l \phi(x_{<j}^i))} \\ &= \sum_{i=1}^n \sum_{j=1}^t W_{x_j^i} \phi(x_{<j}^i) - \sum_{i=1}^n \sum_{j=1}^t \log \sum_{l=1}^V \exp(W_l \phi(x_{<j}^i)) \end{aligned}$$

# (Large) Language Models



$$D = \{X^i\}_{i=1}^n, \quad X^i = \{x_j^i\}_{j=1}^t$$

$$p(X)$$

1. Build probabilistic models
2. Derive loss function (by MLE or MAP....)
3. Select optimizer

Stochastic Gradient Descent



# Summary

- Unification is the path to AGI
  - One model for every task

# Summary

- Unification is the path to AGI
  - One model for every task
- Pretraining of LLM: MLE for unsupervised learning
- RNNs have bottleneck, restricting the flexibility
- Attention is actually flattening RNN
- Transformer is deep attention

Q&A

Presentation Sign-up deadline today!  
Group 7, 11, 18, and 21